# Statistical Methods for Organ Transplant

**A THESIS**
**SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL**
**OF THE UNIVERSITY OF MINNESOTA**
**BY**

**Shannon B. McKearnan**

**IN PARTIAL FULFILLMENT OF THE REQUIREMENTS**
**FOR THE DEGREE OF**
**DOCTOR OF PHILOSOPHY**

**Dr. Julian Wolfson and Dr. David Vock**

**July, 2021**

# Acknowledgements

I want to first thank my advisors Dr. Julian Wolfson and Dr. David Vock, without whom this dissertation would not have been possible. I am so grateful for their support and expertise since my first day as a student at the University of Minnesota. Thank you for always encouraging me to think critically and communicate effectively. Thank you for your support in my development as a researcher and for opening the doors for me to many collaborative opportunities.

Thank you to collaborators on the SMART team Dr. Liz Marai, Dr. Guadalupe Canahuate, and Dr. Dave Clifton for their financial support, data, and contributions to the feature selection for support vector regression methodology portion of this work. Thank you to the Organ Procurement Transplant Network for the use of their incredibly rich dataset on organ transplant that has provided the basis of the data applications throughout this dissertation. I would also like to thank Dr. Erika Helgeson and Dr. Peter Huckfeldt for serving on my committee and providing valuable feedback on my work.

Thank you to Dr. Patrick O'Connor and Dr. Gabriela Vazquez Benitez at HealthPartners Institute for their continued collaboration. Thank you for the valuable feedback and the use of your data for my first publication, which while not published in this work, was an invaluable learning experience and impacted my statistical writing skills

immensely.

Thank you to Dr. Nicole Basta for her mentorship, for setting a wonderful example of leading a lab, and for the opportunity to work on numerous collaborative projects with her and her network. The skills I gained working with Dr. Angela Ulrich, the Canadian Longitudinal Study on Aging team including Dr. Cristina Wolfson, all of the COVID-19 vaccine tracker team, and many of Dr. Nicole Basta's other students and collaborators, are invaluable. While these projects are not published in this work, they have certainly guided me to becoming the biostatistician I am today.

Thank you to all of the Division of Biostatistics faculty and staff for providing an overwhelmingly positive education. In particular, thank you to Sally Olander for her guidance and for going above and beyond in her role to ensure I and other students felt at home in the Division.

My fellow PhD students, Chuyu Deng, Yuan Zhang, Evan Olawsky, Mengli Xiao, Andrew DiLernia, Charles Cain who were by my side throughout this experience and made it truly memorable. Without your friendship and support, I would not be the statistician or the person I am today.

Finally, I would like to thank my family, for whom these thanks are only a small measure of my immense gratitude: to my mom Mary, my dad Paul, and my brother Jack, and my vast extended family. Thank you for always asking about my research and for your loving support throughout my academic career and my life.

## Abstract

In this dissertation, we propose novel statistical methods to improve clinical decision support for organ transplant donors and recipients, using data from the United Network for Organ Sharing national registry. In our first project, we develop a feature selection method for support vector regression in order to benefit from the method's flexibility while combating overfitting. Support vector regression is advantageous due to its use of a kernel for flexibility and computational efficiency; penalized methods for feature selection limit the choice in kernel to finite dimensional transformations and are thus insufficient. We propose a novel feature selection method for support vector regression based on a genetic algorithm that iteratively searches across potential subsets of covariates to find those that yield the best performance according to a user-defined fitness function. We apply our method to predict donor kidney function one year after transplant.

In our second project, we develop an estimator for marginal survival under a dynamic treatment regime for organ transplant, where treatment is defined as the patient's decision to accept or decline an organ when it is offered to them. We apply our method to kidney transplant patients to recommend thresholds of the quality of organ for acceptance. In our third project, we again utilize the genetic algorithm's flexible optimization, this time to identify optimal treatment regimes. We define the treatment regime as a decision list in order to develop our method. We apply our method to identify treatment regimes for liver transplant patients who may wish to undergo a simultaneous kidney transplant. Overall, we develop novel methods in diverse fields of statistics tailored for the organ transplantation context, and we demonstrate their performance and meaningful clinical implications via simulations and real data examples.

# Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

Tools that predict patient outcomes are valuable for healthcare providers, informing them as they make decisions about patient care, including choosing among treatment options or frequency of follow-up visits, and recommending an optimal dosing schedule. Predicting patient outcomes can also be useful in allocating scarce resources, as in organ transplantation. Patients who need an organ transplant must wait until one becomes available that is considered a good match for them; over 113,000 patients are currently waiting for a transplant in the United States. Because the number of patients waiting for an organ exceeds the number of available organs and many physiological factors are involved in a successful transplant, the matching system for allocating potential organs is complex and varies by organ type.

Patients and their healthcare providers face many decisions in the transplantation process; my dissertation work is driven by the over-arching goal to create novel statistical methods to improve clinical decision support in this process. For example, living organ donors may be concerned about the impact of organ transplant to their own health. Improved prediction of their outcomes following transplant will aid in the choice of having the major surgery. In addition, for patients awaiting an organ, accepting the

first organ that is offered by the matching system may not be the best choice; patients must weigh the quality of the organ for their long-term benefit against the likelihood of receiving a better organ in the future. Finally, some patients may benefit from having multiple organs transplanted simultaneously. Patients must weigh the necessity of the additional organ with its potential benefit and regulators must consider the impact of multi-organ transplants on the other patients on the transplant waiting list. The following chapters of this work focus on addressing these questions in diverse fields of statistics by developing statistical methods tailored for the organ transplantation context.

## 1.1 Statistical motivation

### 1.1.1 Feature selection for support vector regression

The utility of predictions of potential transplant benefits relies on the accuracy of such predictions. Machine learning methods including popular methods such as neural networks, support vector machines, and ensemble methods are powerful and flexible tools for prediction. A "flexible" model can more closely adapt to the shape of the data it is supplied with, yielding accurate predictions; however, allowing for too much flexibility in a model can lead to over-fitting, in which the model fits too closely to the exact data it is trained on and thus loses accuracy in its application to outside contexts. This problem is particularly likely to occur when many patient characteristics or features are used. To counteract over-fitting, improve predictive performance, and increase the interpretability of results, the number of patient features used in the model can be narrowed down via feature selection. However, traditional feature selection techniques designed for regression problems, such as LASSO, often do not often generalize to more complex machine learning methods.

As such, in Section 2, we propose a feature selection method for the machine learning method support vector regression (SVR), which extends the support vector machine method used for classification (Drucker *and others*, 1997). SVR utilizes a kernel for flexibility and computational efficiency. The use of a non-linear kernel makes SVR most advantageous when the functional relationship between predictors and outcome is non-linear, especially because the non-linear relationship does not need to be pre-specified for accurate predictions as they would be for linear regression. Common approaches for feature selection, such as penalization, would apply only to the linear kernel and are thus insufficient. To meet this need, we propose a feature selection method for SVR based on the genetic algorithm (Goldberg and Holland, 1988), a general-purpose optimization technique that mimics biological evolution. The genetic algorithm searches widely over the potential solutions, while zeroing in on those that give good results. We show in Section 2 that our proposed method yields higher predictive accuracy than SVR without feature selection and accuracy that is as good as or better than alternate methods such as LASSO or random forest, in most scenarios.

### 1.1.2 Dynamic treatment regimes for organ transplantation

A dynamic treatment regime (DTR) is a decision rule that changes over time based on individualized information about a patient that may change over time, such as the quality of their health. Statistical methods for DTRs have two main aims: 1) to estimate the causal effect of the DTR on patient outcomes and 2) to estimate which treatment regime would be optimal. Current methods to estimate the optimal DTR focus on the effect of the different regimes on a single patient, assuming all other patients make no changes to their behavior. Because of the limited availability of resources for organ transplantation, these existing methods are insufficient; a recommendation to accept only the highest quality of organs is not reasonable to recommend all patients to follow.

Existing methods for estimating effects of DTR also do not account for the uncertainty in when a patient will be offered an organ. To account for this, we define a DTR for organ transplantation in terms of the patient's decline of an offer for a specific type or types of organs. In Section 3, we propose an estimator for the causal effect on survival of a specified DTR for declining one type of organ when it is offered. We extend this estimator to specifically apply to types of organs defined by a continuous quality metric and propose a method to estimate the optimal threshold for the quality of organ to be accepted.

## 1.2 Clinical motivation

### 1.2.1 Living kidney donor outcomes

The development of a feature selection method for support vector regression in Section 2 was motivated in part with the aim to aid potential living kidney organ donors in making an informed decision on whether to perform the transplant. Patients in need of a kidney transplant can benefit greatly from a living donor transplant. While some evidence has shown minimal long-term consequences of kidney donation, other evidence has shown an increased risk of end-stage renal disease among donors (Ibrahim *and others*, 2009). We aim to further investigate the impact of kidney donation on the donor's renal function post-transplant and identify factors that could indicate a donor is at additional risk for future complications. Data for this analysis were collected from January 2015 to December 2019 from the United Network for Organ Sharing national registry. We focus our analysis on patients aged 18 or older who donated kidneys for living donor transplant.

### 1.2.2 Dynamic treatment regimes for kidney transplant

As described in Section 1.1.2, the statistical and clinical motivation for our development of a novel estimator for the causal effect of a dynamic treatment regime are intertwined, as the need for this estimator arises specifically from the lack of available organs for all patients. Because patients do not control when they receive an organ offer, we define a treatment in this context as the patient's decision to accept or decline an organ when it is offered to them. To implement the estimator for marginal survival under a dynamic treatment regime for organ transplant, we first focus solely on kidney transplant patients. Data for this analysis were collected from the Scientific Registry of Transplant Recipients (SRTR). SRTR's database retains records of all transplants, organ donations, and transplant candidates on the waiting list to receive an organ in the United States. The primary source of data in the SRTR database is the Organ Procurement and Transplantation Network (OPTN), which collects data from transplant programs and organ procurement organizations about candidates and the waiting list. The database includes characteristics of patients while they are listed on the waiting list and at the time of transplant, as well as characteristics of the transplant and characteristics of the organ donor. Post-transplant follow-up data on patients is also present in the database where available.

It's important to note that not all organs available for transplant are of equal quality. One reason we selected kidneys as the initial focus for data application of this method is the high-quality metric established for use in quantifying the quality of a kidney available for transplant. The Kidney Donor Profile Index (KDPI) is a metric that summarizes donor factors and compares the quality of the kidney relative to other deceased donor kidneys. KDPI is calculated based on 10 factors about the donor: age, height, weight, ethnicity, history of hypertension, history of diabetes, cause of death, serum creatinine,

hepatitis C virus status, and donation after circulatory death status. All of these factors are readily available with minimal amounts of missingness in the SRTR database. As we will discuss in Section 3, we develop our method for estimating causal effect of a treatment regime when the patient declines organs of a specific type. Estimating the causal effect on survival of treatment regimes that indicate whether patients should decline organs that are low-quality, as measured by KDPI, is relevant in practice.

### 1.2.3 Treatment regimes for simultaneous liver and kidney transplant

Many liver transplant candidates also experience renal dysfunction, which can negatively impact their survival and outcomes following transplant (O'Leary *and others*, 2016). Simultaneous liver kidney (SLK) transplantation has become increasingly common, particularly since the MELD score began to be used for liver allocation in 2002, increasing from 2.5% in 1994 to 10.3% in 2009 (Singal *and others*, 2013). However, for some patients, renal dysfunction may be reversible following improved health after liver transplantation alone. Therefore, identifying patients for whom SLK transplant would be most beneficial is crucial for the allocation of scarce kidneys.

Changes in the Organ Procurement Transplant Network (OPTN)'s allocation policy for SLK are as recent as 2017; limited analysis of the impact of these policies has been conducted. A recent literature review on SLK transplant criteria and outcomes described the gaps in allocation criteria and indicated the need for further examination of how SLK allocation is conducted, particularly for those with acute kidney injury (Singal *and others*, 2019). As such, in Section 4, we apply our method using the genetic algorithm to identify a decision rule for whether or not a given patient listed for liver transplant should undergo liver transplant alone or SLK transplant.

## 1.3 Overview

In the remainder of the dissertation, we discuss statistical developments designed to answer questions of prediction in organ transplantation. In Section 2, we introduce the support vector regression model and discuss the lack of feature selection methods appropriate for it. We propose a feature selection method based on a genetic algorithm and evaluate its results in simulations and in the motivating data example of predicting non-survival lung transplant outcomes. In Section 3, we transition to the causal inference question of dynamic treatment regimes under a limited resource scenario. We propose an estimator of the causal effect of a treatment regime for acceptance of an organ for transplant, addressing the counterfactual setting in which a patient receives an organ. We evaluate its results in simulations and in the application to identify thresholds for the quality of kidney that recipients on the waiting list should accept when offered. In Section 4, we conclude with the proposal of a genetic algorithm-based optimization for treatment regimes defined as decision lists and evaluate its results for identifying when patients should undergo simultaneous liver and kidney transplant versus liver alone transplant.

# Chapter 2

# Feature selection for support vector regression

## 2.1 Summary

Support vector regression (SVR) is particularly beneficial when the outcome and predictors are non-linearly related. However, when many covariates are available, the method's flexibility can lead to overfitting and an overall loss in predictive accuracy. To overcome this drawback, we develop a feature selection method for SVR based on a genetic algorithm that iteratively searches across potential subsets of covariates to find those that yield the best performance according to a user-defined fitness function. We evaluate the performance of our feature selection method for SVR, comparing it to alternate methods including LASSO and random forest, in a simulation study. We find that our method yields higher predictive accuracy than SVR without feature selection. Our method outperforms LASSO when the relationship between covariates and outcome is non-linear. Random forest performs equivalently to our method in some scenarios, but more poorly when covariates are correlated. We apply our method to predict donor

kidney function one year after transplant using data from the United Network for Organ Sharing national registry.

## 2.2 Introduction

In this section, we propose a feature selection technique for support vector regression (SVR) (Drucker *and others*, 1997), a machine learning technique for predicting continuous outcomes that derives from the popular support vector machine (SVM) method for classification. In a clinical setting, it is important to predict outcomes accurately, while not requiring a number of predictors that would be unreasonable to collect. In a retrospective study, researchers may also be interested in using feature selection to identify key relationships between predictors and the outcome of interest. We first develop our method in a general context to demonstrate its viability and accuracy, then we apply it to a national registry of kidney transplant donors to predict kidney function following transplant.

SVR utilizes a kernel for flexibility and computational efficiency. The use of a non-linear kernel makes SVR most advantageous when the functional relationship between predictors and outcome is non-linear, especially because the non-linear relationship does not need to be pre-specified for accurate predictions as they would be for linear regression. Penalized methods for feature selection limit the choice in kernel to finite dimensional transformations and are thus insufficient. While most feature selection methods for SVM have primarily focused on classification without adaptation to SVR, some are able to be easily adapted and others have been proposed for SVR. For example, kernel iterative feature extraction (known as KNIFE), which includes a penalty term on weighted features within a kernel, and kernel-penalized SVM (KP-SVM) have been shown to be successful feature selection methods in the classification setting (Allen,

2013; Maldonado *and others*, 2011). Recursive feature elimination (RFE) has been demonstrated to correctly select features for various kernel problems, including SVR specifically (Dasgupta *and others*, 2019; Dasgupta and Huang, 2020). Other proposed methods for conducting feature selection in the regression case require pre-specification of the number of features to select for the model (Yang and Ong, 2011) or increase the tuning required with added SVR hyper-parameters (Maldonado and Weber, 2010).

Our approach to SVR feature selection involves a genetic algorithm, an optimization technique modeled after evolutionary processes. With the genetic algorithm, candidate sets of selected variables are "bred" together and passed on over subsequent generations, with "strongest" (i.e., most relevant) set of potential covariates retained for each generation. Genetic recombination operators such as crossover and mutation are applied to maintain some diversity in each generation (Goldberg and Holland, 1988; Leardi *and others*, 1992; Yang and Honavar, 1998). Genetic algorithms have been used in combination with support vector machines for classification problems in areas such as brain MRI classification, blood cell recognition, and microarray-based tumor classification (Kharrat *and others*, 2010; Osowski *and others*, 2009; Peng *and others*, 2003; Li *and others*, 2005). Various methods have jointly used both SVR and the genetic algorithm, most commonly to optimize the parameters of the SVR model (Liu *and others*, 2013; Saravanan and Sailakshmi, 2015; Wu *and others*, 2009). We are not aware of any prior work that tackles the problem of SVR feature selection using the genetic algorithm.

The rest of Section 2 is organized as follows. In Section 2.3, we briefly review the basics of SVR and genetic algorithms and describe our proposed method for using the genetic algorithm to perform feature selection for SVR. In Section 2.4, we compare it to other methods for feature selection in a simulation study. In Section 2.5, we illustrate the proposed method in an application to a national registry of kidney transplant donors. We discuss the results and provide concluding remarks in Section 2.6.

## 2.3 Methods

### 2.3.1 Support vector regression

Support vector regression is an extension of the support vector machine method for classification which constructs an optimal separating hyperplane between classes (Cortes and Vapnik, 1995). We will briefly review the SVR methodology; Smola and Scholkopf (2004) provide a more detailed explanation.

Consider data features $\mathbf{X}_i$ in the space $\mathcal{X} \in \mathbb{R}^L$ along with continuous outcome $Y_i$ for individuals $i = 1, \ldots, n$. One of the advantages of the SVR algorithm is its ability to capture non-linear relationships. This is implemented in the algorithm by pre-processing the input data $\mathbf{X}_i$ with a transformation to some new feature space. We can write this transformation as the function $\Phi(\boldsymbol{x}) : \mathcal{X} \mapsto \mathcal{H}$, where $\mathcal{H}$ is a high-dimensional, possibly infinite-dimensional, reproducing kernel Hilbert space. Define the function to approximate the relationship between $Y_i$ and $\boldsymbol{X}_i$ in the feature space as an inner product between a weight vector $\boldsymbol{w}$ and the transformed predictors $\Phi(\boldsymbol{x})$ as

$$f(\boldsymbol{x}) = \langle \boldsymbol{w}, \Phi(\boldsymbol{x}) \rangle_{\mathcal{H}} + b. \tag{2.1}$$

Then, the support vector regression problem consists of optimizing

$$\min_{\boldsymbol{w},b} \left[ \frac{1}{2} \|\boldsymbol{w}\|_{\mathcal{H}}^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*) \right] \tag{2.2}$$

where $C$ is a positive-valued constant regularization parameter and $\xi_i, \xi_i^*$ are slack variables constraining the margin of error. Instead of applying the same loss function as for a support vector machine, we use an $\epsilon$-insensitive loss function Drucker *and others*

(1997) defined as

$$L(y, f(x))_\epsilon = \begin{cases} 0 & \text{if } |y_i - f(x_i))| < \epsilon \\ |y_i - f(x_i)| - \epsilon & \text{otherwise} \end{cases} \tag{2.3}$$

This loss function allows for some errors by treating errors that are small enough, as defined by $\epsilon$, simply as zero, while measuring the loss of larger errors by their magnitude. Minimizing Equation 2.2 under this loss function corresponds to the following constraints on the optimization:

$$\begin{aligned} f(x_i) - y_i &\leq \epsilon + \xi_i \\ y_i - f(x_i) &\leq \epsilon + \xi_i^* \\ \xi_i, \xi_i^* &\geq 0, \forall i = 1, \ldots, n. \end{aligned} \tag{2.4}$$

Note that we can avoid direct computation and specification of the transformation $\Phi(\mathbf{X})$ by applying a kernel function, defined as the inner product $k(\mathbf{x}, \mathbf{x}') = \langle \Phi(\mathbf{x}), \Phi(\mathbf{x}') \rangle_{\mathcal{H}}$. Because we have defined $\mathcal{H}$ as a reproducing kernel Hilbert space, we must select a kernel that meets the Mercer conditions, which includes commonly used kernels such as the Gaussian radial basis function. In an alternate formulation of the SVR optimization problem known as the "Lagrangian dual formulation," the problem can be rewritten such that the transformation $\Phi(\boldsymbol{x})$ is only involved via dot products. This formulation allows for the computational use of a kernel function instead; because the transformation $\Phi(\boldsymbol{x})$ is typically high-dimensional or infinite-dimensional in order to capture the non-linearity of the function, it would be computationally intensive to directly compute, and a kernel function is more feasible. A quadratic programming algorithm is then applied to solve the dual formulation optimization problem. In our simulations and data application, we apply a radial basis function kernel to account for

lack of prior knowledge about the relationships between the data and outcome, rather than assuming a linear or quadratic relationship (two other commonly used kernels). This kernel is defined as $k(\mathbf{x}, \mathbf{x}') = \exp(-\gamma||\mathbf{x} - \mathbf{x}'||^2)$. $\gamma$ is an additional parameter introduced in the selected kernel function and is jointly tuned via grid search across potential values with the cost parameter $C$.

Many common approaches to feature selection involve adding a $L_1$ penalty term to the loss function and inducing sparsity (Tibshirani, 1996; Zou, 2006; Han *and others*, 2015). While this could be applied to the SVR model with a linear kernel, it is not applicable in general. Many of SVR's strengths are derived from its handling of non-linear data using alternative kernels; thus, a feature selection technique that is applicable for any kernel choice is preferred. Descent-based methods are not easily applicable for this question, as we cannot define the changes in the SVR model based on the inclusion or exclusion of a certain covariate when it is optimized in the dual formulation, which we seek to use for its kernel properties.

### 2.3.2 Genetic algorithms and feature selection

The genetic algorithm is a general-purpose optimization technique designed to model biological evolutionary practices in which only the fittest individuals reproduce and pass on their genetic information to the following generation, leading to stronger individuals as generations pass. Genetic operators such as mutation and crossover are used to increase diversity of genetic information across generations. Genetic algorithms are well-suited to problems where the objective function implies a complex solution space that cannot be easily traversed by descent-based methods. The genetic algorithm is a flexible technique and has been utilized in many fields of study for a wide range of problems, including but not limited to clustering analyses, multiobjective optimization, and multiprocessor scheduling (Maulik and Bandyopadhyay, 2000; Horn *and others*,

1994; Hou *and others*, 1994).

The first step of the genetic algorithm consists of defining a procedure for encoding the objects to be optimized as binary strings, e.g. 011010100. Each string represents a "chromosome", i.e. a distinct object, and each binary component is referred to as a "gene". The population is initialized with a number of such chromosomes via random generation. Individuals are then "bred" over multiple generations using various genetic operators such as recombination and mutation. The potential solutions as represented by the chromosomes are evaluated at each generation according to the objective (fitness) function; the "fittest" offspring are retained at each generation. The algorithm ends after a given number of generations or when there is no more diversity in the population of potential solutions. The final solution has the best performance, as determined by the fitness function, among the last generation of chromosomes.

Though not originally designed for feature selection, when the genetic algorithm is adapted to feature selection, the encoding process is straightforward; the binary components of each chromosome are indicators of whether each potential covariate is included or excluded in the statistical model.

The various genetic operators applied are elitism, selection, crossover, and mutation. The generations of chromosomes following the initialization are created via elitism, in which the top-performing portion is automatically advanced to the next generation, and selection, in which some chromosomes are selected to form the "reproducing population". Different selection schemes can be applied. In linear-rank selection, chromosomes are selected with replacement for the reproducing population with probability linearly assigned based on rank of performance according to the designated fitness function. In roulette wheel selection, chromosomes are selected with probability proportionate to the value of their fitness function. Blickle and Thiele (1995) describe several other selection schemes.

Crossover is applied to the reproducing population of chromosomes, with the intent to create diversity in the possible solutions for the next generation. The reproducing population chromosomes are randomly matched to form pairs referred to as "parents". Information is swapped, i.e. crossover is applied, between the two individuals with a fixed crossover probability. For pairs undergoing crossover, information is swapped between the two parents to form two new chromosomes referred to as "offspring," which replace the parents in the new population. For pairs not undergoing crossover, parent chromosomes are treated as the offspring. A variety of crossover methods can be applied. In single-point crossover, modeled after traditional biological crossover between chromosomes, a randomly selected gene is chosen within the chromosome as the changeover point; then, each offspring receives the genetic material from one parent prior to that point and the other parent after that point. In uniform crossover, information is exchanged at the gene level rather than dividing the chromosome into two segments (Syswerda, 1989). Each gene is exchanged between the two parents with a fixed probability. Additional genetic diversity is then induced via mutation. Each gene within each chromosome is randomly modified by simply changing the indicator of whether or not the particular variable is included in analysis; mutation of a particular gene is performed with a fixed probability that is typically small. Following the iterative application of these genetic operators and the assessment of chromosomes under the fitness function, the final solution indicates a set of covariates to be included for analysis in the prediction method.

### 2.3.3 Feature selection for support vector regression model

Unlike other methods such as generalized linear models, SVR does not have an obvious way to organize and explore the model space. Applying the genetic algorithm can therefore add functionality not otherwise available. Our feature selection technique for

the SVR model based on the genetic algorithm is displayed graphically in Figure 2.1 and can be described by the following steps:

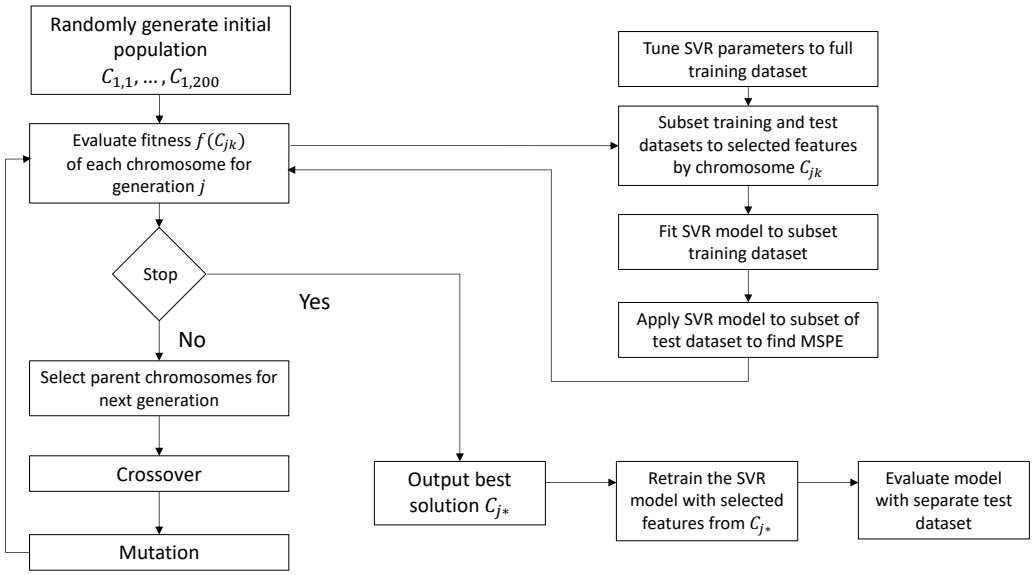


Figure 2.1: Flowchart detailing the feature selection method for SVR using the genetic algorithm.

1. For features $\boldsymbol{X} \in \mathbb{R}^L$, define a chromosome $C_{jk}$ composed of L genes $g_{jkl} \in \{0, 1\}$, where each entry $g_{jkl}$ indicates whether or not covariate $l$ is selected for inclusion in the model. For $k = 1, \ldots, K$ chromosomes, initialize the first generation, $j = 1$, of $K$ chromosomes via random generation.

2. Tune the SVR hyper-parameters cost and $\gamma$ by grid search, using the training dataset with all features.

3. Evaluate each chromosome $C_{jk}$ under a fitness function $f(C_{jk})$ designed to evaluate the model given the selected features indicated by the $g_{jkl}$ entries of the chromosome.

    i. Subset training and test datasets to selected features by $C_{jk}$.

ii. Fit SVR model to subsetted training data using optimal hyper-parameters as found in step 2.

iii. Apply SVR model to test dataset and calculate $f(C_{jk})$.

4. Generate the reproducing population of chromosomes for the following generation.

   i. Elitism: Automatically include the top-performing chromosomes as evaluated by the fitness function in the following generation. Include the top $K * p_E$ chromosomes, where $p_E$ is the proportion of elitism.

   ii. Selection: Select $K * (1 - p_E)$ high-performing chromosomes for inclusion in the reproducing population, using linear-rank selection. Exclude the $K * p_E$ chromosomes from Step 4ii from the selection process.

5. Apply genetic operators to the $K * (1 - p_E)$ chromosomes generated in step 4ii to form the next generation.

   i. Crossover: Randomly pair the reproducing population chromosomes to form $K*(1-p_E)/2$ pairs referred to as "parents". Apply uniform crossover between the two individuals with probability $p_C$.

   ii. Mutation: Randomly modify a small number of $g_{jkl}$ to induce genetic variation. Change the indicator of each 0/1 $g_{jkl}$ in each chromosome in the new population with probability $p_M$.

6. Repeat steps 3-5 for $j = 1, \ldots, N$ iterations, leading to a final generation of chromosomes. Select covariates for inclusion in the model per the chromosome with the best performance according to the fitness function $f$, denoted $C_{N*}$.

Elitism proportion for use in selection was set to $p_E = 0.05$, a commonly used value. Linear-rank selection was chosen as the selection schematic, in order to maintain

diversity in potential solutions in the case that fitness function values vary in magnitude. We applied uniform crossover with a crossover probability $p_C = 0.5$, to eliminate possible dependence on the order in which the variables are coded in the chromosome stemming from single-point crossover. Mutation probability was set to $p_M = 0.1$. Mutation probability is often kept small; the goal is to induce some additional genetic diversity, not completely change the potential solutions.

A population size of $K = 200$ was chosen and the algorithm was repeated for $N = 200$ generations. The genetic algorithm utilizes a fitness function at each iteration to evaluate the performance of each individual. We compared versions of the genetic algorithm implementing mean squared prediction error (MSPE) and Bayesian information criterion (BIC) as fitness functions. Separate training and test data sets were used to calculate the value of the fitness function within the iteration of the genetic algorithm.

## 2.4 Simulation study

### 2.4.1 Simulation methods

We assessed the performance of our feature selection method for support vector regression with a set of simulations. Since our proposed method combines feature selection (via the genetic algorithm) with flexible regression modeling, we designed our simulation study to investigate the impact of both these elements on prediction accuracy under a wide variety of scenarios. To isolate the effects of feature selection, we compared our method to an SVR without selection. To isolate the effects of the regression modeling, we combined genetic algorithm feature selection with a linear regression model, denoted LR. We compared our method to the RFE method for SVR, using a change-point analysis to determine the stopping point for feature elimination. We also considered

alternative prediction approaches such as LASSO-based variable selection with a linear regression model, as well as a random forest. Average percentage of simulations in which at least one related covariate is selected, average number of related covariates included, average percentage of simulations in which at least one unrelated covariate is selected, average number of additional unrelated covariates included, and average mean squared prediction error (MSPE) are reported to assess the selection of variables and model performance. Performance of the selected set of predictors was evaluated using a separately generated test data set. 500 simulations were performed for each scenario. The following different data generating scenarios were considered.

For scenario 1, uncorrelated normal random covariates were generated with underlying associated quadratic terms associated with the outcome. Covariates $\boldsymbol{X}_C = (X_1, \ldots, X_p)$ were generated independently from a normal distribution with mean zero, variance one. Quadratic terms were added to the design matrix
$\boldsymbol{X} = [X_1^2 \; X_2^2 \; \ldots X_{\pi p}^2 \; X_1 \; X_2 \; \ldots X_p]$, where $\pi \in [0, 1]$ was varied across scenarios to change the proportion of covariates related to the outcome. We let $z_i$ be the $i^{th}$ row of $\boldsymbol{X}$; the outcome $\boldsymbol{Y}$ was generated as $Y_i = \boldsymbol{\beta}' z_i + \epsilon_i$ with $\epsilon_i \sim \text{Normal}(0, 1)$. We set $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3)$, where $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are vectors each of length $p * \pi$ corresponding to the quadratic and linear predictors associated with the outcome and $\boldsymbol{\beta}_3$ is a vector of length $p * (1 - \pi)$ set to be zero in all scenarios. The values of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ were varied across simulations to adjust the amount of relative weight of the quadratic term and its related linear term on the outcome respectively, for the following scenarios: positive values of $\boldsymbol{\beta}_1$ only, indicating that the relationship between covariates and outcome is purely quadratic; positive values of both $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, but with greater values for one of the two, shifting from a more quadratic to a more linear relationship; positive values of $\boldsymbol{\beta}_2$ only, indicating that the relationship is only linear. Following the data generation in this manner, only covariates $\boldsymbol{X}_C$ were supplied to the prediction methods in order to

assess how well the methods would capture the quadratic effects in their predictions and how often the associated main effect terms would be selected via the feature selection processes.

For scenario 2, uncorrelated normal random covariates were generated with underlying associated interaction terms associated with the outcome. Covariates $\boldsymbol{X}_C = (X_1, \ldots, X_p)$ were generated independently from a normal distribution with mean zero, variance one. As in scenario 1, we let $z_i$ be the $i^{th}$ row of $\boldsymbol{X}$ and the outcome $\boldsymbol{Y}$ was generated as $Y_i = \boldsymbol{\beta}' z_i + \epsilon_i$ with $\epsilon_i \sim \text{Normal}(0, 1)$. Pairwise interaction terms were added to the design matrix for all covariates with corresponding main effect terms, such that $\boldsymbol{X} = [X_1X_2 \ X_1X_3 \ \ldots X_{\pi p-1}X_{\pi p} \ X_1 \ X_2 \ \ldots X_p]$. We set $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3)$, where $\boldsymbol{\beta}_1$ is a vector of length $\binom{\pi p}{2}$ corresponding to the interaction terms, $\boldsymbol{\beta}_2$ is a vector of length $p * \pi$ corresponding to the linear terms, and $\boldsymbol{\beta}_3$ is a vector of length $p * (1 - \pi)$ set to zero. As with the quadratic data generation scenario, interaction terms were not supplied to the methods for use in feature selection or prediction.

For scenario 3, correlated multivariate normal data was generated with varying levels of association with the outcome. The $N \times p$ design matrix $\boldsymbol{X}$ was formed by partitioning into equal-sized blocks of correlated covariates. Each block was generated from a multivariate normal distribution with mean zero and first order autoregressive correlation structure with $\rho = 0.5$. Given $z_i$ the $i^{th}$ row of $\boldsymbol{X}$, the outcome $\boldsymbol{Y}$ was generated as $Y_i = \boldsymbol{\beta}' z_i + \epsilon_i$ with $\epsilon_i \sim \text{Normal}(0, 1)$. The value of $\boldsymbol{\beta}$ was defined to match the block design of the covariates in two different ways: one such that the main effect terms were correlated with irrelevant terms but uncorrelated with each other, and one such that the main effect terms were correlated with each other but not with irrelevant terms. Technical details of the settings for each scenario, including values of $\boldsymbol{\beta}$, are presented in Section 1 of the Appendix.

### 2.4.2 Simulation results

In scenario 1, we assessed the performance of our method in the presence of underlying quadratic relationships between the outcome and covariates, while only providing the corresponding linear covariates to the model. Results are displayed in Figure 2.2 and Figure 2.3. Tables displaying the full results are presented in Section 2 of the Appendix. Our feature selection method for SVR selected all of the relevant covariates in both the 50 and 300 covariate scenarios, where the proportion of relevant covariates $\pi$ is equal to 0.1 and 0.05 respectively, regardless of the level of quadratic relationship. On average, the BIC fitness function in the genetic algorithm yielded the selection of fewer extraneous, unrelated variables than the use of the MSE fitness function, in all scenarios. In addition, the accuracy of predictions from our method, as measured by MSPE, was lower in all scenarios than that of the SVR model implemented without any feature selection. While the LASSO method selected the correct variables when the relationship was primarily linear, it did not do so when the relationship shifted to quadratic. The same pattern was seen for the genetic algorithm feature selection method applied to linear regression models. The random forest method maintained roughly the same accuracy of predictions no matter the level of quadratic relationship between outcome and predictors; additionally, no feature selection was implemented with the random forest. The RFE method performed similarly to random forest in the 50 covariate scenario, however, it does include feature selection and it correctly selected predictors. In the 300 covariate scenario, random forest outperformed our method in terms of MSPE when the relationship between outcome and predictors was mostly quadratic but still maintained a linear component. In all other cases, our method performed better. While RFE and the GA applied to linear regression selected some of the correct predictors in the 300 covariate scenario, the proportion declined as the

relationship between outcome and predictors became more quadratic.

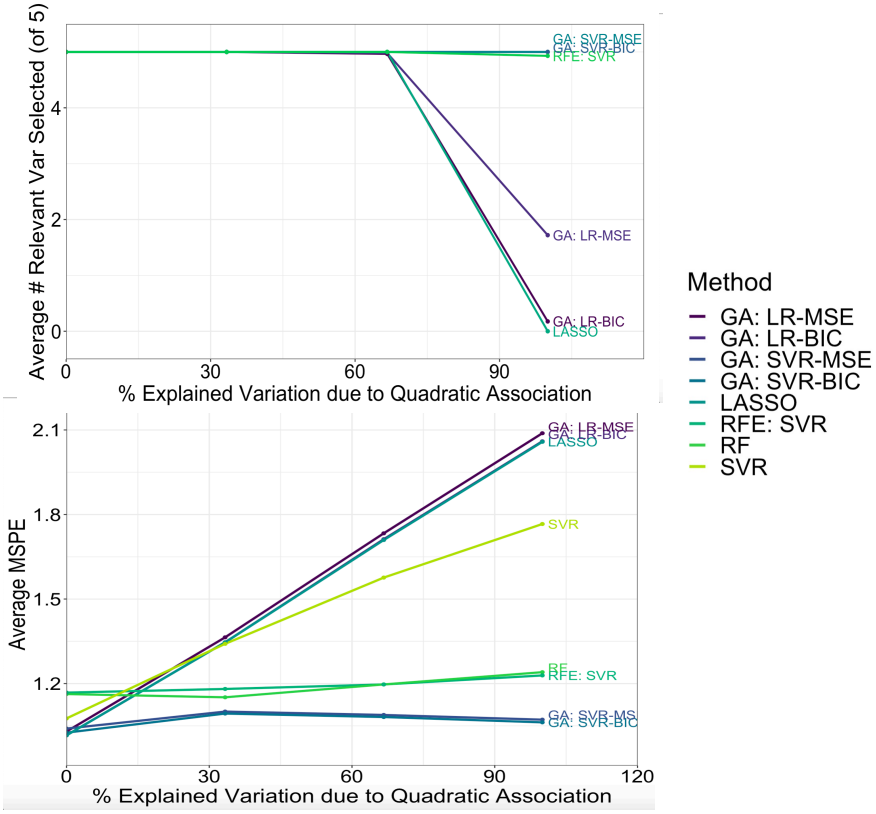


Figure 2.2: Results of simulation 1 setting with 50 total covariates $\mathbf{X}$, 5 of them truly associated with the outcome $\mathbf{Y}$. Trend in average MSPE and number of variables selected by various prediction methods as the amount of true variation in $\mathbf{Y}$ due to $\mathbf{X}$ stemming from the quadratic covariates is increased. Random forest method is not displayed for number of variables due to lack of feature selection.

In scenario 2, we assessed the performance of our method when there are true interactive effects between covariates, while only providing the corresponding linear covariates to the model. Results are displayed in Table 2.1. In this scenario with 50 covariates, when 10 underlying interactive terms were not supplied to the prediction method but 5 related linear terms were, our method for SVR with either the MSE or BIC fitness function used in the genetic algorithm and RFE with SVR yielded the lowest MSPEs. While

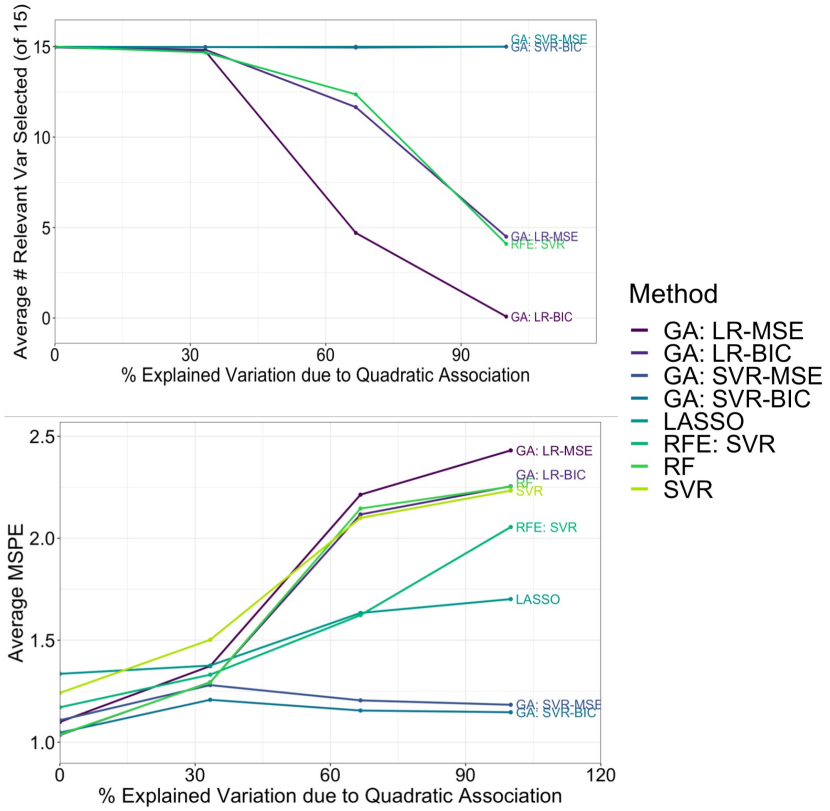


Figure 2.3: Results of simulation 1 setting with 300 total covariates $\mathbf{X}$, 15 of them truly associated with the outcome $\mathbf{Y}$. Trend in average MSPE and number of variables selected by various prediction methods as the amount of true variation in $\mathbf{Y}$ due to $\mathbf{X}$ stemming from the quadratic covariates is increased. Random forest method is not displayed for number of variables due to lack of feature selection.

the random forest method predictions were nearly as accurate, the linear model (using either LASSO or the genetic algorithm for feature selection) and the SVR model without any feature selection left over twice as much of the variance due to X unexplained on average. In addition, our method with the BIC fitness function used in the genetic algorithm selected the fewest extraneous variables for inclusion in the model. The genetic algorithm using BIC in combination with linear regression and LASSO methods

also picked up very minimal extraneous variables. All methods correctly included the covariates that were truly associated with the outcome.

| Feature Selection | Model Fit | % Selection Relevant (1+) | Avg # Relevant Var | % Other Var (1+) | Avg # Additional Var | Avg % Variation Unexplained by Pred. [1] |
|---|---|---|---|---|---|---|
| GA: MSE | LR | 100 | 5 | 100 | 16.12 | 55.36 |
| GA: BIC | LR | 100 | 5 | 6.0 | 0.07 | 53.26 |
| **GA: MSE** | **SVR** | 100 | 5 | 95.2 | 4.30 | 24.31 |
| **GA: BIC** | **SVR** | 100 | 5 | 27.4 | 0.30 | 22.06 |
| RFE | SVR | 100 | 5 | 100 | 3 | 22.76 |
| LASSO | LR | 100 | 5 | 0.6 | 0.12 | 53.26 |
| None | RF | 100 | 5 | 100 | 45.0 | 27.47 |
| None | SVR | 100 | 5 | 100 | 45.0 | 51.22 |

Table 2.1: Results of simulation 2 setting with 50 total covariates $\mathbf{X}$, 5 of them truly associated with the outcome $\mathbf{Y}$, and 10 pairwise underlying interaction effects not supplied to the model for selection or prediction. Columns represent the proportion of simulations in which at least one of the relevant covariates is selected, the number of relevant covariates selected on average (out of 5), the proportion of simulations in which at least one unassociated covariate is selected, the number of unassociated covariates selected on average, and a scaled MSPE described in [1].

[1] Column displays average across simulation iterations of the proportion of true variation in $\mathbf{Y}$ due to $\mathbf{X}$ that is left unexplained by the prediction. For example, in the case where the total variance in $\mathbf{Y}$ is 2, where half is due to the error term and half is due to the covariates $\mathbf{X}$, a MSPE of 1.05 would be reported as 5% in this column. Of the explainable variance 1, the model misses 5%.

In scenario 3, we compared the performance of our method to competing methods when the covariates are correlated. Results are displayed in Table 2.2 and Table 2.3. 50 covariates were used in this scenario, comprised of equal-sized correlated blocks of 10 covariates. We applied two different implementations of correlated data: one in which truly related covariates were correlated with extraneous variables and one in which truly related covariates were correlated with each other. In both cases, all methods

correctly selected the relevant variables for inclusion in the model. Similar trends in both the accuracy of the predictions and the variables selected were seen in both cases. The implementation of our feature selection method for SVR yielded a sizable decrease in MSPE compared to SVR without any feature selection. However, in this scenario, LASSO and the linear regression with a genetic algorithm and BIC fitness function methods yielded both the lowest MSPEs and the fewest amounts of extraneous variables selected for inclusion in the model. As seen in the previous scenarios, the use of the BIC fitness function in the genetic algorithm as opposed to the MSE fitness function led to far fewer additional variables being selected. In addition, the random forest and RFE methods yielded predictions less accurate than any other method.

| Feature Selection | Model Fit | % Selection Relevant (1+) | Avg # Relevant Var | % Other Var (1+) | Avg # Additional Var | Avg % Variation Unexplained by Pred. [1] |
|---|---|---|---|---|---|---|
| GA: MSE | LR | 100 | 5 | 100 | 15.2 | 2.96 |
| GA: BIC | LR | 100 | 5 | 7.8 | 0.08 | 1.62 |
| **GA: MSE** | **SVR** | 100 | 5 | 100 | 12.39 | 3.87 |
| **GA: BIC** | **SVR** | 100 | 5 | 15.0 | 0.16 | 2.53 |
| RFE | SVR | 100 | 5 | 100 | 3 | 15.63 |
| LASSO | LR | 100 | 5 | 100 | 0.12 | 1.6 |
| None | RF | 100 | 5 | 100 | 45.0 | 14.55 |
| None | SVR | 100 | 5 | 100 | 45.0 | 7.17 |

Table 2.2: Results of simulation 3 setting with 50 total covariates $\mathbf{X}$, 5 of them truly associated with the outcome $\mathbf{Y}$, and each correlated to 9 variables unrelated to the outcome with AR-1 correlation structure.

[1] Column displays average across simulation iterations of the proportion of true variation in $\mathbf{Y}$ due to $\mathbf{X}$ that is left unexplained by the prediction. For example, in the case where the total variance in $\mathbf{Y}$ is 2, where half is due to the error term and half is due to the covariates $\mathbf{X}$, a MSPE of 1.05 would be reported as 5% in this column. Of the explainable variance 1, the model misses 5%.

In the supplementary material, we present additional simulation results for alternative signal-to-noise ratios and an alternate choice in kernel for SVR. The overall trends

| Feature Selection | Model Fit | % Selection Relevant (1+) | Avg # Relevant Var | % Other Var (1+) | Avg # Additional Var | Avg % Variation Unexplained by Pred. [1] |
|---|---|---|---|---|---|---|
| GA: MSE | LR | 100 | 10 | 100 | 13.47 | 3.23 |
| GA: BIC | LR | 100 | 10 | 0 | 0 | 2.07 |
| **GA: MSE** | **SVR** | 100 | 10 | 100 | 11.30 | 3.57 |
| **GA: BIC** | **SVR** | 100 | 10 | 0.6 | 0.01 | 2.39 |
| RFE | SVR | 100 | 7.99 | 0 | 0 | 31.38 |
| LASSO | LR | 100 | 10 | 0.02 | 2.0 | 1.96 |
| None | RF | 100 | 10 | 100 | 40.0 | 21.97 |
| None | SVR | 100 | 10 | 100 | 40.0 | 7.31 |

Table 2.3: Results of simulation 3 setting with 50 total covariates $\mathbf{X}$, 10 of them truly associated with the outcome $\mathbf{Y}$ and correlated with each other with AR-1 block correlation structure.

[1] Column displays average across simulation iterations of the proportion of true variation in $\mathbf{Y}$ due to $\mathbf{X}$ that is left unexplained by the prediction. For example, in the case where the total variance in $\mathbf{Y}$ is 2, where half is due to the error term and half is due to the covariates $\mathbf{X}$, a MSPE of 1.05 would be reported as 5% in this column. Of the explainable variance 1, the model misses 5%.

in the results presented here are similar to those seen in the lower signal-to-noise ratio scenario.

## 2.5 Application to kidney transplant registry data

As first introduced in Section 1.2.1, data for this analysis were collected from January 2015 to December 2019 from the United Network for Organ Sharing national registry. We focus our analysis on patients aged 18 or older who donated kidneys for living donor transplant. 21,121 patients are included in the analysis and 32 covariates are available regarding donor demographics, quality of the transplanted organ, and donor health. Missing covariate data were imputed using Multivariate Imputation by Chained Equations (MICE). Descriptive characteristics and imputation details are presented in

Section 4 of the Appendix. We consider the outcome estimated glomerular filtration rate (eGFR) one year post-transplant; eGFR is a measure of kidney function based on the patient's creatinine, age, and gender. We evaluate performance of the methods by splitting the data into separate training, evaluation (for use in evaluating the genetic algorithm fitness function), and test data sets. We use 60% of the data for training data, 20% for evaluation data, and 20% for testing data.

Results of the application of our method and comparative methods to predict eGFR at 1 year following kidney transplant are displayed in Table 2.4. We find that using our feature selection method for SVR with an MSE fitness function yields a lower mean squared prediction error (MSPE) than all tested alternatives except random forest, most significantly improving performance over SVR without any feature selection.

The genetic algorithm with a BIC fitness function heavily penalizes the number of variables selected in this application, selecting only one variable, pre-transplant eGFR, to be included in the model for both linear regression or SVR. The genetic algorithm with a MSE fitness function selects 15 variables for inclusion in the linear regression and 13 variables for inclusion in the SVR model, though there is some variation in the variables selected. In addition to the donor's pre-transplant eGFR, both methods select donor blood type, education history, race, and BMI. RFE used with SVR selects similar variables: pre-transplant eGFR, race, blood type, and BMI.

These results are not inconsistent with the results of the simulation studies we performed, in which our methods had lower MSPE than SVR without feature selection in all scenarios examined. Some improvement in MSPE was seen in our method when using the MSE fitness function, which allowed for more variables to be included in a richer model. In methods with fewer variables selected, the same few were seen in common, indicating consistency across the methods in identifying predictive variables of post-transplant eGFR. In addition, these results illustrate the impact of different

fitness criterion. In this example, because of the strong relationship between one specific predictor and the outcome, the BIC fitness function selects only the one key predictor and misses additional predictors that are beneficial to the model. This contrasts to our simulation results, in which relevant predictors had equal impact and the results generally showed a lower MSPE for the BIC fitness criterion than the MSE fitness criterion. Analysts applying our method should be mindful of the structure of their own data when choosing the fitness criterion.

| Feature Selection | Model Fit | No. Var. Selected | MSPE |
|---|---|---|---|
| GA: MSE | LR | 15 | 112.6 |
| GA: BIC | LR | 1 | 118.5 |
| **GA: MSE** | **SVR** | 13 | 110.3 |
| **GA: BIC** | **SVR** | 1 | 114.2 |
| RFE | SVR | 5 | 120.1 |
| LASSO | LR | 2 | 114.5 |
| None | RF | 32 | 104.1 |
| None | SVR | 32 | 122.6 |

Table 2.4: Performance results of GA-SVR method and comparison methods applied to predict eGFR for living transplant kidney donors at 1 year post-transplant.

These results are not inconsistent with the results of the simulation studies we performed, in which our methods had lower MSPE than SVR without feature selection in all scenarios examined. Random forest performed similarly to our method in multiple scenarios involving non-linear relationships between outcome and covariates, which we see here as well. In this setting, one variable provided a large amount of predictive power; in settings where more variables are related to the outcome, our method may provide a greater benefit over alternatives.

## 2.6 Discussion

The biggest gain offered by support vector regression compared to more traditional statistical methods is its well-handling of non-linear data without the need to pre-specify transformed or interaction terms; however, the same structure of SVR that yields this benefit impedes the development of typical feature selection methods. To overcome this disadvantage, we developed a genetic algorithm based approach that allows for ease of continued use of SVR's various kernel functions. The computation of non-linear SVR does not directly depend on the dimensionality of the data, but rather the number of samples; while some may take this to mean that SVR is inherently resistant to typical problems of dimensionality, we have not found that to be the case. Our GA-SVR feature selection method yields better predictive accuracy than SVR without feature selection in all simulation experiments examined as well as in our data application, indicating otherwise. Random forest, an alternative machine learning technique also popular for its predictive accuracy, performs comparably to our method in many settings of our simulation setting, though not out-performing it. Notably, our method out-performs random forest most significantly in the case of correlated predictors. This is also a scenario in which the LASSO method is known to have poor performance and indicates a situation in which use of our method may be most advantageous. In addition, we found that our method yielded considerable improvement in predictive accuracy over LASSO when covariates had quadratic or interactive effects.

As with all uses of support vector regression, proper tuning of parameters and choice in kernel function are necessary steps for good performance. Choice in kernel is discussed extensively in existing literature. We conducted a sensitivity analysis for our simulation study using a linear kernel instead of a radial basis kernel for SVR; results are presented in Section 5 of the Appendix. As expected, our method with a linear kernel performed

well when the data generating mechanism was linear, and poorly when it was not. Given that the use of the radial basis kernel when the data generation was linear was comparable to alternate methods, we suggest using the radial basis kernel in cases when the data generation is either unknown or known to be non-linear. A further examination of additional alternate kernels may be of interest. Alternative methods for tuning of parameters have been developed; it may be of interest in the future to examine their impact in conjunction with our feature selection method. Computational effort for our method is not insurmountable especially with the implementation of stopping criteria; however, due to the iterative nature of the genetic algorithm, if computational speed is a major concern an alternative such as random forest may be preferable. A further examination of alternative fitness functions used in the genetic algorithm and their impact on feature selection is also warranted.

# Chapter 3

# Estimating the causal effect of a dynamic treatment regime in organ transplantation

## 3.1 Summary

In standard clinical contexts, a dynamic treatment regime (DTR) is a decision rule that indicates at what time a patient should begin treatment, based on their health information. In the context of organ transplantation, a DTR under this definition would indicate at what time a patient should receive an organ for transplant. However, since the number of patients awaiting a transplant greatly outnumbers the number of organs available, such a DTR would be infeasible in practice, as the patient is not in control of when they will be offered an organ from the allocation system. Not all organs available for transplant are of the same quality; a patient offered an organ must decide whether it would benefit them more to receive the organ offered or stay on the waiting list, where they may later receive a higher-quality organ offer. As such, we define one

class of regimes in which a patient is instructed which types of organs to avoid if offered, where type of organ could indicate the quality of the organ or other characteristics. We propose an estimator of the marginal survival probability a patient would have under the given DTR, using inverse probability weighting. We propose an extension of this estimator in the case of a continuous metric of organ quality that optimizes the threshold of the metric for the treatment regime. We present simulation results evaluating the performance of our method and demonstrating its success against a naive method. We also present results applying our method to kidney transplantation using data from the Scientific Registry of Transplant Recipients.

## 3.2 Introduction

A dynamic treatment regime (DTR) is a clinical tool which dictates a patient's treatment or intervention to be assigned based on their prior treatment and covariate history (Moodie *and others*, 2007). A DTR both personalizes a patient's treatment and allows for changes in treatment over time, as the patient's characteristics may change and as prior treatment may impact future treatment. As mentioned in Section 1.1.2, to successfully implement a DTR in practice, one has two main questions: 1) what is the causal effect of the DTR on patient outcomes? and 2) based on this causal effect, which treatment regime should be applied (i.e., which is optimal)?. Methods for estimating the causal effect on survival of a dynamic treatment regime and later, methods for estimating the optimal dynamic treatment regime, have been studied in detail. Common approaches to estimating the causal effect of a DTR on survival while adjusting for time-varying confounders include Robins *and others* (1992)'s g-estimation of the parameters of a structural nested accelerated failure time models and Hernán *and others* (2000)'s use of inverse probability weighting to estimate the parameters of a marginal structural

Cox proportional hazards model. Several authors have compared these methods in real-data applications and raise disadvantages of these methods, such as the implications of artificially censoring data (Joffe *and others*, 2012; Westreich *and others*, 2012).

The majority of literature surrounding DTRs focuses on treatments that are abundantly available. We can refer to these treatment regimes as *deterministic*; if the regime declares that a patient should receive the treatment at a given time, the patient can without question receive the treatment. For patients awaiting an organ transplant, the treatment, in this case the receipt of an organ, is not abundantly available. As previously mentioned, the allocation system for organ transplant is complex, with many more patients awaiting an organ than organs available. As such, if a DTR recommended that a patient "begin treatment," i.e. receive an organ, on a specific day, the patient's ability to implement the treatment depends on whether they are offered an organ at that time. We refer to this treatment as *probabilistic*. Instead of defining the DTR in terms of the time at which a patient should receive treatment, as would be the case for a deterministic regime, we instead define the DTR in terms of the patient's choice to accept or decline an organ of a specific quality when it is offered to them.

A natural conclusion might be that if we could model the allocation process step by step, we could estimate the causal effect of a DTR using this information on when patients receive and accept organ offers. Numerous statistical methods have been developed to model a patient's choice in accepting or declining an organ offer with the goal of using this information to inform organ allocation policy using methods such as Markov decision processes and queuing theory (Kim *and others*, 2015; Su *and others*, 2004; Alagoz *and others*, 2007). To model the entire allocation process requires several models with various assumptions on each; each step of the process needs to be modelled, from patient entry to the waiting list, changes in patient status as they remain on the waiting list, organ arrivals and organ offers, and patient acceptance of organs.

The Scientific Registry of Transplant Recipients (SRTR) uses this approach in their development of simulated allocation models. Because this approach requires correct specification of the models for each step of the allocation process, we prefer to avoid this method. Schaubel *and others* (2006) developed a method that does not involve simulating the allocation process, in which they apply sequential stratification to estimate the survival benefit of a accepting a kidney from an expanded criteria donor, i.e. a kidney likelier to have worse outcomes due to certain health characteristics of the donor, compared to remaining on the waitlist with the potential to receive a traditional, lower-risk kidney in the future. However, their method is specific to the extended donor criterion for kidneys used to categorize kidneys as high-risk. In our method, we propose more broadly defining the treatment regimes of interest by not relying on one specific metric. We define the treatment regime as declining a given type of organ, with the type specified by the user; for example, we could assess the treatment regime of declining all organ offers from donors aged 50 or over.

Standard methods to estimate the effect of a DTR focus on the effect of the different regimes on a single patient, assuming that all other patients make no changes to their behavior. The organ allocation system and waiting list that control when patients receive an organ offer mean that this is not a valid assumption in this context. Whether or not a patient with a higher priority accepts or declines an organ will impact a patient with lower priority's treatment. In recent work, Boatman and Vock (2018) developed an estimator of patient survival when following a given DTR for organ acceptance; their method incorporates whether or not all other patients follow the same strategy, addressing this concern. As expected, they do find that the effect of a DTR on patient survival varies if all or none of the other patients also follow that strategy. Here, we propose a more flexible approach that more fully captures the setting in which a patient is offered an organ. We consider the counterfactual "organ setting" of a patient, which

may include characteristics such as their region or blood type, for which allocation policies can vary and change over time. Rather than modelling the entire allocation process, we instead aim to account for the frequency at which a patient is offered an organ. We build inverse probability weights for the patient's probability of following their observed treatment and covariate history under the counterfactual organ setting and specified DTR, using these to estimate the marginal probability of survival.

As mentioned, we define the treatment regime as the patient's acceptance of an organ of a specific quality. When this quality can be described by a continuous metric, we further extend our method to identify the optimal threshold of the metric for accepting or declining an organ. We use our estimator to estimate the marginal probability of survival under a grid of thresholds.

The rest of Section 3 is organized as follows. In Section 3.3.1, we describe our proposed method for estimating the causal effect of a specified DTR for what type of organs a patient should decline. In Section 3.3.2, we describe the extension of our method to estimate the optimal threshold for accepting or declining an organ when the metric of organ quality is continuous. In Section 3.4, we describe the implementation of this method for kidney transplant data obtained from the Scientific Registry of Transplant Recipients and its results. In Section 3.5, we describe the simulation study designed to examine the performance of our metric and its results. In Section 3.6, we summarize our work and discuss avenues of improvement and extension for this method.

## 3.3 Methods

### 3.3.1 Proposed estimator for survival under a dynamic treatment regime

Suppose that we have a population of $n$ patients on the organ transplant waiting list, each of whom who we follow for up to $m$ days after they join the waiting list, during

which time they may or may not receive a transplant and they may or may not die. Let $i$ indicate the patient, $i = 1, \ldots, n$, and $j$ indicate the day since the patient was added to the organ waiting list.

These two events, receipt of an organ transplant and death, are both of interest to the patient. In this scenario, receipt of an organ transplant is the treatment involved in the dynamic treatment regime, while survival is the key outcome of interest. As such, we differentiate the following variables with the superscript $O$ relating to the organ transplant and the superscript $P$ relating to the patient's survival. Let $\mathbf{L}_{ij}^O$ be characteristics for the $i$-th patient on the $j$-th day after listing which are associated with time-to-transplant but are not associated with survival, such as the patient's geographic region or blood type. Let $\mathbf{L}_{ij}^P$ be characteristics for the $i$-th patient on the $j$-th day after listing which are associated with survival; note that these characteristics could be associated with both time-to-transplant and survival, such as the patient's age. Then, let $\mathbf{L}_{ij} = (\mathbf{L}_{ij}^O, \mathbf{L}_{ij}^P)$. We use the overbar notation to denote history, where $\bar{\mathbf{L}}_{ij}$ is then the history of covariates for patient $i$ through $j$ days after listing.

Let $N_{ij}^P$ be an indicator for the death of the $i$-th patient on the $j$-th day after listing and let $Y_{ij}^P$ be an indicator for the whether or not the $i$-th patient is at risk of death on the $j$-th day after listing. Let $N_{ij}^O$ be an indicator for the $i$-th patient's receipt of an organ transplant on the $j$-th day after listing and let $Y_{ij}^O$ be an indicator for whether or not the $i$-th patient is at risk of receiving an organ transplant on the $j$-th day after listing; note that we restrict analysis to patients who have not received a prior organ transplant, such that $Y_{ij}^O$ will equal 0 on all days following a patient's transplant.

As previously mentioned, not only is the receipt of an organ transplant important to patients, so is the quality of that transplanted organ. Suppose that there are $K$ different types of organs, for example, different levels of Kidney Donor Profile Index as introduced in Section 1.2.2. We further differentiate the above indicator for organ

transplant by each $K$ type of organ, letting $N_{ij}^{O,k}$ be an indicator for the $i$-th patient's receipt of an organ transplant of type $k$ on the $j$-th day after listing.

expit We can estimate the cause-specific discrete-time hazard of receiving an organ of type $k$ from a deceased donor while on the waiting list by fitting any model for a binary outcome. We fit this model with outcome $N_{ij}^{O,k}$, recall this is the indicator for $i$-th subject receiving an organ transplant of type $k$ on the $j$-th day after listing, and predictors $\mathbf{L}_{ij}^{O}$ and $\mathbf{L}_{ij}^{P}$ among those for whom $Y_{ij}^{O} = 1$, or those at risk of receiving an organ transplant. To illustrate this, assume that these models are logistic regression models, though alternate transformations could be used. We approach the logistic regression by stacking the data into a person-period data set for all days a patient is eligible for transplant and then fitting the following model on the stacked data set:

$$\log\Big(\frac{\lambda^{N}(t_j|\mathbf{L}_{ij})}{1-\lambda^{N}(t_j|\mathbf{L}_{ij})}\Big) = \log\Big(\frac{\lambda_0^{N}(t_j)}{1-\lambda_0^{N}(t_j)}\Big) + \gamma_{k,O}^{T}\mathbf{L}_{ij}^{O} + \gamma_{k,P}^{T}\mathbf{L}_{ij}^{P}, \tag{3.1}$$

where $\lambda^{N}(t_j)$ is the hazard of receiving an organ transplant at time $j$ and $\lambda_0^{N}(t_j)$ is the baseline hazard. Then, $P(N_{ij}^{O,k} = 1|Y_{ij}^{O} = 1, \mathbf{L}_{ij}) = 1/\big(1+\exp\big\{-\gamma_{k,O}^{T}\mathbf{L}_{ij}^{O} - \gamma_{k,P}^{T}\mathbf{L}_{ij}^{P}\big\}\big)$, the probability of patient $i$ receiving an organ of type $k$ on their $j$-th day after joining the waiting list, given that they are at-risk for transplant on that day and given their covariate history until day $j$.

We can then find the probability of a patient $i$ following their observed treatment history, i.e. if and when they received an organ and of what type, over the course of $m$ days, for all $K$ types of organs, given their covariate history, in the observed organ setting in which patients do not make any changes to when they accept or reject organ offers.

This probability of a patient $i$ following their observed treatment history $(n_1^{O,k}, \ldots, n_m^{O,k})$

for $k = 1, \ldots, K$ and $(y_1^O, \ldots, y_m^O)$ given covariate history $(\boldsymbol{\ell}_1, \ldots, \boldsymbol{\ell}_m)$ is given by

$$\prod_{j=1}^{m} \prod_{k=1}^{K} \left\{ \pi^{O,k}(\ell_j)^{n_j^{O,k}} \left(1 - \pi^{O,k}(\ell_j)\right)^{1-n_j^{O,k}} \right\}^{y_j^O}, \tag{3.2}$$

where $\pi^{O,k}(\boldsymbol{\ell}_j) = P(N_{ij}^{O,k} = 1 | Y_{ij}^O = 1, \mathbf{L}_{ij} = \boldsymbol{\ell}_j; \gamma_{k,O}, \gamma_{k,P})$, the probability estimated by the logistic regression model above. We can estimate the above quantity by substituting in the estimated parameters from the logistic regression model (i.e., $\hat{\gamma}_{k,O}, \hat{\gamma}_{k,P}$). We drop the $i$ subscripts in the above equation for interpretability, though the probability refers to a single patient $i$. Note that we use the $O$ superscript here for both $Y_{ij}$ and $N_{ij}$, as we are referring to the *organ* transplant or treatment of a subject, not the *patient* survival outcome.

Now, we consider what would have happened had the patient followed a specific treatment regime for when to decline an organ offer; this rule is supplied by the researcher to answer a question of clinical relevance, for example, we may want to know what would happen if patients are instructed to decline all organ offers of the lowest-quality organs. Many potential treatment regimes could be developed; however, we focus first on this style of treatment regimes. Thus, assume that in the counterfactual world a patient follows a particular regime such that they decline organs in group $k^*$. Patients could decline organs in multiple categories of the $K$ types of organs, but for now we focus on the case in which only one type is included in $k^*$. Additionally, let $\mathbf{L}_{ij}^O$ be set to $\boldsymbol{\ell}_{ij}^{O*}$, the characteristics that do not affect transplant outcomes and are only related to time-to-transplant may vary in the counterfactual organ setting. This is important as we would expect that the rate at which patients receive organ offers would likely change under the proposed treatment regime; this would be reflected here.

Then, the probability of a patient $i$ following their observed treatment history in

this counterfactual organ setting under the selected treatment regime is given by

$$\prod_{j=1}^{m}\left[\left[\prod_{k\neq k^*}\left\{\pi^{O,k}(\boldsymbol{\ell}_j^P,\boldsymbol{\ell}_j^{O*})^{n_j^{O,k}}\left(1-\pi^{O,k}(\boldsymbol{\ell}_j^P,\boldsymbol{\ell}_j^{O*})\right)^{(1-n_j^{O,k})}\right\}\right]\times\left[\left(1-n_j^{O,k^*}\right)\right]\right]^{y_j^O}. \quad (3.3)$$

The first term in Equation 3.3 is similar to the product above in Equation 3.2, however, we consider only the $k$ organs that the patient would be willing to accept excluding organ type $k^*$ and the probabilities of organ receipt are conditional on the counterfactual organ setting covariates $\boldsymbol{\ell}_j^{O*}$, rather than the observed $\boldsymbol{\ell}_j^O$. The second term indicates that for patients who received an organ that under the selected treatment regime they would have rejected, we set the probability of following their observed treatment history in the counterfactual setting to zero.

Utilizing these two probabilities to form inverse probability weights, we can estimate the discrete-time hazard of death $\lambda_m^{k^*}$ if a patient $i$ were to avoid organs of type $k^*$ and was in an organ environment $\boldsymbol{\ell}_{im}^{O*}$ over a period of $m$ days. Define Equation 3.2 as $\pi(\bar{N}_{im}^{O,k},\bar{Y}_{im}^O,\bar{\mathbf{L}}_{im})$ and Equation 3.3 as $\pi_{k^*,l_{im}^{O*}}(\bar{N}_{im}^{O,k},\bar{Y}_{im}^O,\bar{\mathbf{L}}_{im})$. Certain standard causal assumptions are required to relate the observed data and the potential outcomes: positivity, which states that there is a non-zero probability of a patient following a particular treatment regime, consistency, which states that the observed outcome is equal to the potential outcome under the treatment actually received, and sequential ignorability, which states that the probability a patient receives an organ transplant does not depend on future potential outcomes (conditioned on the patient characteristics $\mathbf{L}_{ij}^P$ and $\mathbf{L}_{ij}^O$). Then, we can estimate $\lambda_m^{k^*}$ by solving the following weighted estimating equations

$$\sum_{i=1}^{n}\frac{\pi_{k^*,l_{im}^{O*}}\left(\bar{N}_{im}^{O,k},\bar{Y}_{im}^O,\bar{\mathbf{L}}_{im}\right)}{\pi\left(\bar{N}_{im}^{O,k},\bar{Y}_{im}^O,\bar{\mathbf{L}}_{im}\right)}\left\{N_{im}^P-\lambda_m^{k^*}Y_{im}^P\right\}=0. \quad (3.4)$$

Under the causal assumptions described above, (3.4) is a mean-zero estimating function and, thus, under standard regularity conditions, the estimator is consistent and asymptotically normal. For $m = 1, \ldots, t$, this yields the $t$-dimensional estimate of discrete-time hazard of death $\hat{\boldsymbol{\lambda}}^{k^*} = \{\lambda_1^{k^*}, \ldots, \lambda_t^{k^*}\}^T$. Then, the marginal survival probability $\hat{S}^{k^*}$ under the counterfactual treatment regime and organ setting through day $t$ can be estimated as $\hat{S}^{k^*} = \prod_{m=1}^{t}\{1 - \hat{\lambda}_m^{k^*}\}$. With this estimator, we successfully avoid modeling each individual step of the organ allocation process.

### 3.3.2 Identifying the threshold for organ acceptance

When the type of organs $k$ can be described by a continuous metric, we can use our estimator for marginal survival to estimate the threshold of that metric that yields the highest marginal survival at a specified time $\tau$. For example, the Kidney Donor Profile Index for kidney transplants used in our data application example is a continuous metric that measures the quality of the donor kidney.

As previously defined, there are $K$ different types of organs. In the case where the type of organ is described by a continuous metric $Q$, where $Q \in \mathbb{R}$ and a high-value of $Q$ is associated with a lower quality of organ, we define $q_k$ as the value of the metric for organ type $k$. Then, rather than defining $k^*$ as avoiding only organs of type $k$, we define $k^*$ as avoiding any organs where the value of the metric is $q_k$ or higher. We suggest the use of a large range of possible values of $Q$, such as $K = 100$ or higher.

For each value of $K$, we solve Equation 3 at day $\tau$ to get $\hat{S}^{k^*}$. Then, we smooth these results by fitting a spline regression model $G(k)$, using b-splines of degree 10 for the thresholds $q_k$. We take the maximum $\hat{G}(k)$ to find the maximum smoothed estimate of our marginal survival. Thus, the value of $q_k$ at this maximum is recommended as a threshold, such that individuals should accept organs only of values $q_1, \ldots, q_{k-1}$ in order to maximize survival at time $\tau$.

## 3.4 Application to kidney transplant

### 3.4.1 Methods

As first introduced in Section 1.2.2, data for this analysis were collected from the Scientific Registry of Transplant Recipients (SRTR). Because both the allocation of organs and potential outcomes vary between adult and pediatric patients, we limit analysis to adults aged 18 years or older when they first join the waiting list. Organ allocation methods also function differently for patients who are receiving multiple transplants simultaneously; we limit analysis to patients receiving only a kidney transplant. For similar reasons, we also limit analysis to patients who have not received a prior kidney transplant. We include all patients who joined the waiting list in 2000 or later. In total, 435,751 patients are included in the analysis, with 208,874 (47.9%) of them receiving a deceased-donor transplant prior to March 3, 2018 (the end of follow-up for the iteration of the SRTR database used in this analysis).

As an additional consideration, kidney transplant patients are in the unique situation of potentially receiving living donor organs; unlike most organs which cannot be donated while a donor is still living, because most people have two functioning kidneys, one of them can be donated while the donor is still alive. While some other types of organs are possible to receive from a living donor, such as the liver, kidney transplants are by far the most frequent living donor transplantation. For simplicity of analysis, we treat patients who receive a living kidney donation as we would any other patient no longer at risk of receiving an organ transplant.

As introduced in Section 1.2.2, the Kidney Donor Profile Index (KDPI) is a metric measuring the quality of a deceased donor kidney recovered for transplantation. Higher scores of KDPI are associated with lower quality of the organ; for example, a KDPI of

85%, a commonly referenced value for a "high" score, indicates that the kidney has a relative risk of graft failure greater than 85% of other recovered kidneys. KDPI is an interpretable metric that maps Kidney Donor Risk Index (KDRI) onto a cumulative percentage scale. KDRI gives the relative increase in the hazard of graft loss compared to the median donor in a given year. For example, a KDRI value of 1.4 indicates that the kidney has a hazard of graft loss 1.4 times the hazard of the typical (median) donor kidney. We use KDRI as values of $Q$ in the statistical framework presented in the previous section to distinguish between types of organs that a patient may want to decline. The reference population used in calculation of KDRI is all kidneys recovered for transplantation in the prior calendar year. Though values of KDRI are not included in the SRTR database, instructions for calculation of the scores and scaling to the reference population are published by the Organ Procurement and Transplantation Network each year; all clinical and demographic donor factors required for the calculation of KDRI are present in the SRTR database with minimal amounts of missingness. We use the 2010 coefficients for KDRI calculation and scaled to the 2010 reference population.

We use logistic discrete-time hazard models to estimate $\pi^{O,k}(\mathbf{l}_j) = P(N_{ij}^{O,k} = 1|Y_{ij}^{O} = 1, \mathbf{L}_{ij} = \mathbf{l}_j)$. As previously defined, this is the probability of a patient $i$ following their observed treatment history $(n_1^{O,k}, \ldots, n_m^{O,k})$ for $k = 1, \ldots, K$ and $(y_1^O, \ldots, y_m^O)$ given covariate history $(\mathbf{l}_1, \ldots, \mathbf{l}_m)$. This probability is used in the calculation of the inverse probability weights introduced in Section 3.3. We use the discrete-time unit of one month and calculate these hazard models separately for each of the 58 organ procurement organizations (OPO). Because we are using continuous values of KDRI as the $K$ types of organs, we simplify calculation of the logistic discrete-time hazard models to 4 groupings of $K$, corresponding to bins of KDPI percentiles of 0-29, 30-59, 60-84, 85-100. We use covariates of candidate age, blood type, diabetes status, height, race, time on dialysis, and calculated panel reactive antibody (cPRA) and include interaction

terms for race and gender, age and time on dialysis, diabetes status and time on dialysis.

We then apply Equations 1-3 from Section 3.3 to yield estimates of $\hat{S}^{k^*}$ for $K = 100$ values of KDRI, ranging from 0.8023 to 1.62, which correspond to KDPI values of 30% and 90% respectively. We compare our method to a naive estimator of survival in which any patient accepting an organ with KDRI greater than or equal to $q_k$ is censored. As described in Section 3.3.2, we estimate the threshold for organ acceptance to maximize estimated survival at 2 years and at 5 years.

### 3.4.2 Results

Table 3.1 provides descriptive characteristics of the 435,751 patients included in our analysis for the patient characteristics used in fitting discrete time hazard models to form our inverse probability weights. Table 3.2 displays descriptive characteristics of the 143,377 donors whose kidneys were transplanted for the donor characteristics used in calculation of KDRI and KDPI.

Figures 3.1 and 3.2 display a comparison of our smoothed estimate of survival and a naive estimate of survival at each threshold of KDRI for declining an organ. Recall that the threshold indicates that patients should accept a kidney that is offered to them that is under the threshold. In both figures and at all thresholds, we estimate survival to be much higher than the naive estimate.

In addition, we identified the thresholds that maximize our smoothed estimate of marginal survival at 2 and 5 years. We found that a KDRI of 1.396 and 1.447 maximized 2-year and 5-year respectively, corresponding to KDPI values of 81% and 84% respectively. From a clinical standpoint, it is reassuring that these thresholds are very similar, indicating that short- and long-term survival can both be maximized by declining organs of a similar quality.

| **Patient Characteristics** | N (%) / Mean (sd) |
|---|---|
| Gender | |
| Female | 169,343 (38.9) |
| Male | 266,408 (61.1) |
| Race | |
| Asian | 29,280 (6.7) |
| Black | 127,329 (29.2) |
| Other | 8,452 (1.9) |
| White | 270,690 (62.1) |
| Blood type | |
| A | 137,230 (31.5) |
| AB | 16,025 (3.7) |
| B | 64,820 (14.9) |
| O | 212,799 (48.8) |
| Other | 4,877 (1.1) |
| Diabetes diagnosis | |
| No | 251,644 (57.7) |
| Type I | 15,970 (3.7) |
| Type II | 126,963 (29.1) |
| Other/Unknown | 41,174 (9.4)) |
| Age (years) | 51.75 (13.0) |
| Height (cm) | 170.42 (10.57) |
| Time on dialysis (days) | 18.31 (26.79) |
| Calculated panel reactive antibody (cPRA) | 0.10 (0.24) |

Table 3.1: The mean and standard deviation (if continuous) or frequency and proportion (if categorical) of each covariate are presented for the 435,751 patients used in the analysis. Characteristics presented are those used in the logistic discrete-time hazard models to build inverse probability weights.

## 3.5 Simulation study

We assessed the performance of our estimator for survival under a dynamic treatment regime using a simulation study. We compared our method to a naive estimator that censors any patient accepting an organ that does not fall under that treatment regime. We design our simulation study in the same context as our applied example, kidney

| **Donor Characteristics** | N (%) / Mean (sd) |
|---|---|
| Race | |
| Black | 19,720 (13.8) |
| Other | 123,657 (86.2) |
| History of hypertension | |
| Yes | 39,366 (27.5) |
| No | 103,040 (71.9) |
| Unknown | 971 (0.7) |
| History of diabetes | |
| Yes | 9,988 (7.0) |
| No | 132,690 (92.5) |
| Unknown 699 | (0.5) |
| Hepatitis C virus (HCV) status | |
| Yes | 4,485 (3.1) |
| No | 138,892 (96.9) |
| Meets donation after cardiac death (DCD) criteria | |
| Yes | 19,959 (13.9) |
| No | 123,418 (86.1) |
| Age (years) | 38.65 (16.44) |
| Height (cm) | 168.74 (18.30) |
| Weight (kg) | 79.20 (24.39) |
| Serum creatinine (mg/dL) | 1.16 (0.98) |

Table 3.2: The mean and standard deviation (if continuous) or frequency and proportion (if categorical) of each covariate are presented for the 143,377 donors used in the analysis. Characteristics and selected categories presented are those used in the calculation of KDRI and KDPI.

transplant. As in our applied example, we define the categories of organs for acceptance by the continuous metric KDRI.

To generate the data for our simulations, we begin by treating arrivals to the waiting list as fixed based on our observed data described in Section 1.2.2. We sample with replacement from the covariate histories $(\mathbf{l}_1, \ldots, \mathbf{l}_j)$ in our observed data. We then generate the patient treatment history $(n_1^{O,k}, \ldots, n_j^{O,k})$ and survival $(y_1^O, \ldots, y_j^O)$ for $k = 1, \ldots, K$ levels of KDRI, $j = 1, \ldots, \tau$ days and $i = 1, \ldots, N$ patients. Using the

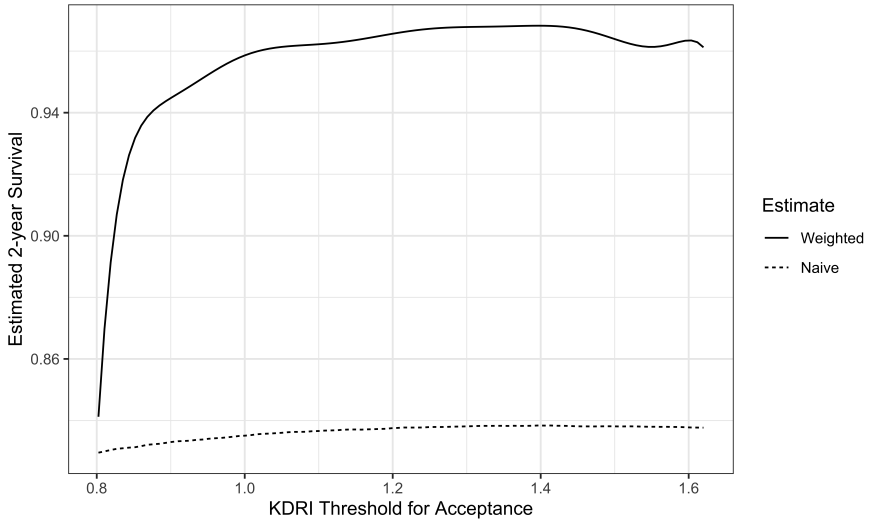


Figure 3.1: Comparison of our smoothed estimate of survival (solid line) and a naive estimate (dotted line) at 2 years at varying KDRI thresholds for declining an organ.

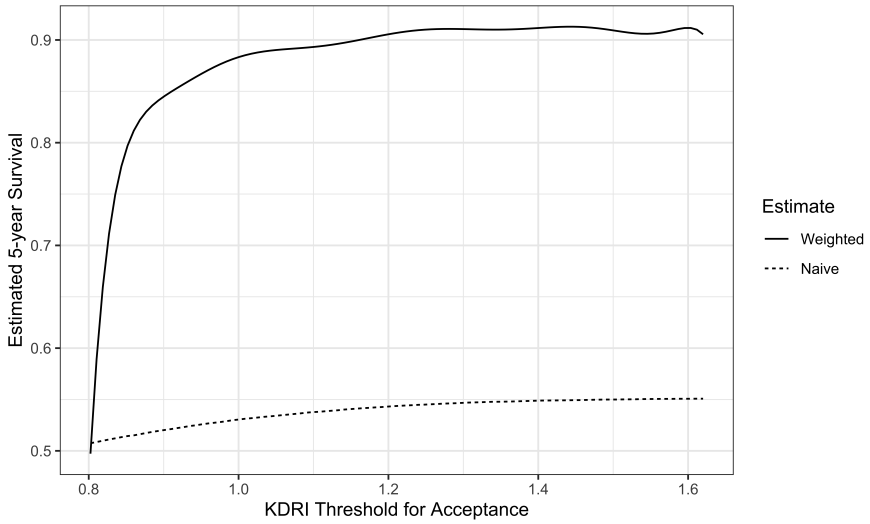


Figure 3.2: Comparison of our estimate of survival (solid line) and a naive estimate (dotted line) at 5 years at varying KDRI thresholds for declining an organ.

estimated parameters $\hat{\gamma}_k, O, \hat{}_k, P$ from Equation 3.1, we generate the probabilities of receiving a transplant at time $t$ to get $n_j^{O,k}$, until $n_j^{O,k} = 1$ for any value of $k$ or the end of follow-up.

We estimated the survival distribution assuming patients follow the treatment regime $k^*$, to decline any organs with KDRI value higher than $q_k$. We use our inverse probability weighted estimator $\hat{S}^{k^*}(t)$ from Section 3 as well as a naive estimator $\hat{S}^N(t)$ that censors individuals when they become non-compliant from the regime, without any weights. While the naive estimator is often used in practice, it does not consistently estimate any causal effect of interest.

The true survival probability $S^{k^*}(t)$ $t$ days after entering the waiting list and following the regime to avoid acceptance of any organs with KDRI higher than $q_k$ is not available in closed form. We estimated the survival curves via Monte Carlo simulation, where the Monte Carlo datasets are independent of the ones used to evaluate the proposed estimator. In the simulated dataset, all patients followed the identified treatment regime and received a survival benefit from transplant if they accepted a kidney.

For 3 levels of KDRI, we sampled 1000 Monte Carlo datasets of 20,000 patients over a 5-year follow-up period. We report the true survival probabilities, bias of the estimators, and coverage probability of 95% confidence intervals for the 3 treatment regimes of interest at two time points. These results are presented in Table 3.3, with KDRI displayed as KDPI for interpretability. We found that bias was larger at all time points for the naive estimator, which consistently underestimated the true survival under the treatment regime. Bias tended to be larger at the 5-year timepoint than the 2-year timepoint for the naive estimator. The simulation results demonstrate that bias was small and coverage probabilities were large for our estimator, indicating good performance. Results did not vary greatly across different thresholds of KDRI/KDPI for organ acceptance; this is a reassuring result, as we would hope to find that our estimator yields good results regardless of the treatment regime of interest.

| KDPI Threshold | $t$ (months) | True Survival | Bias $\hat{S}^N(t)$ | Bias $S^{\hat{k^*}}(t)$ | Coverage Probability |
|---|---|---|---|---|---|
| 30 | 24 | 0.800 | -0.036 | -0.002 | 0.936 |
| 30 | 60 | 0.581 | -0.115 | -0.001 | 0.941 |
| 60 | 24 | 0.824 | -0.048 | 0.002 | 0.939 |
| 60 | 60 | 0.627 | -0.178 | 0.003 | 0.932 |
| 85 | 24 | 0.831 | -0.044 | 0.003 | 0.943 |
| 85 | 60 | 0.640 | -0.201 | 0.002 | 0.930 |

Table 3.3: Simulation results for our estimator and a naive estimator of survival at $t$ in months. $q_k$ KDRI value is mapped to the KDPI threshold for organ acceptance for interpretability. Coverage probability is presented as 95% CIs from simulation.

## 3.6 Discussion

We demonstrated how we can avoid modelling the entirety of the organ allocation process by introducing an estimator that uses inverse probability weights to account for the frequency a patient is offered an organ. Our estimator performs well in simulations when compared to a naive approach that treats patients as censored when they become non-compliant with the treatment regime of interest. Our method relies on defining the type of organs to be avoided in a treatment regime: these types can be categorical or continuous; in the continuous setting, we extend our method to identify the optimal threshold of the metric defining organ quality at which to accept or decline an organ.

We found for kidney transplant with categories of organs defined by organ quality as described by KDPI/KDRI, avoiding kidneys with values of KDPI higher than 81% maximized survival at 2-years after joining the waiting list. Recall that a KDPI of 81% indicates that a donor has higher expected risk of graft failure than 81% of all kidney donors in the reference population (we used donors from 2010 in our analysis). These KDPI thresholds can provide additional information for patients and their providers as they consider whether or not to undergo a transplant. We focused our application in

this work on kidney transplants; examining treatment regimes for other organs would be of interest in future work.

Kidney donation is unique in that organs can also be donated by living donors, given that most people have two functioning kidneys with only one of them being required. While living donation is seen occasionally with other organs such as livers, it is most often seen with kidneys. Expanding the definition of our estimator of marginal survival to handle competing events, such as the receipt of a living donor kidney, is of practical interest.

Additionally, the current work is designed to estimate the probability of survival. Alternate outcomes could be of clinical interest, such as the number of dialysis-free days following kidney transplant. Both survival and quality-of-life are major factors in a patient's choice to receive a transplant; future work may consider the extension of the method to alternate outcomes of interest, which we would expect to vary by type of organ transplant.

# Chapter 4

# Identifying optimal treatment regimes using a genetic algorithm

## 4.1 Summary

While much research has been conducted in identifying treatment regimes that are optimal, often times the treatment regimes are not interpretable by clinical researchers. Treatment regimes that can be represented as a set of decision rules are more interpretable and therefore more useful in practice. We propose an optimization method based on the genetic algorithm for treatment regimes defined as decision rules. We take advantage of the genetic algorithm's ability to search widely across a problem space while narrowing in on the best-performing solution and its adaptability for many types of problems. We apply our method to identify treatment regimes for patients deciding whether to undergo simultaneous liver and kidney transplant or liver transplant alone using data from the United Network for Organ Sharing national registry. We find that without restricting the population level of simultaneous kidney and liver

transplant, nearly all patients with pre-existing renal dysfunction will benefit from simultaneous kidney and liver transplant. Due to the nature of the resource scarcity in kidney transplant, we also identify treatment regimes under various restrictions on the population-wide amount of transplant.

## 4.2 Introduction

A primary goal of precision or personalized medicine is to develop guidelines for treatment decisions to improve the expected outcome of patient's. More formally, an individual treatment rule or treatment regime is a function which maps from a patient's measured characteristics to a set of actions, interventions, or treatments to be undertaken. A substantial amount of statistical research has been conducted in developing methods to identify treatment regimes that are optimal. Many of these methods, such as outcome-weighted learning (OWL), aim to maximize the marginal mean of the outcome for a certain class of treatment regimes (Zhao *and others*, 2012). Augmented inverse probability weighting estimator-based optimization methods have been developed for single time-point and multi-time-point treatment regimes, referred to as C-learning (Zhang *and others*, 2012; Zhang and Zhang, 2018). Fu *and others* (2016) suggest a search algorithm for subgroup identification motivated by OWL. However, the treatment regimes developed by these approaches are not always interpretable by clinicians as they can be composed of linear combinations of many variables.

Feature selection is one tool that is often used to provide more interpretable results; however, combining feature selection with the goal of optimizing a treatment regime presents a number of challenges. One approach to optimizing a treatment regime is to estimate the conditional means and then take the maximum (Moodie *and others*, 2014). The models for the conditional mean are trying to both induce an optimal decision rule

and correct for confounding; adding variable selection could thus lead to confounded estimates. Additionally, using a flexible method for the regression, such as random forest, is useful in this optimization; however, flexible methods do not always allow for straightforward variable selection. Alternatively, one could take the approach of directly maximizing the value of a treatment regime (Zhang *and others*, 2012). Attempting to directly maximize the value of a treatment regime can lead to challenging optimization because the estimated value is often non-smooth in certain parameters; this is only more challenging when also attempting variable selection.

One common way of operationalizing an interpretable treatment regime is through a decision list (Shortreed *and others*, 2011; Laber and Zhao, 2015; Zhang *and others*, 2015). A decision list is a sequence of "if-then" clauses that classify a patient to a recommended treatment based on their characteristics. For example, we could possibly find a simple treatment regime for SLK transplant as follows: if a patient has been on dialysis for longer than 10 days or has been in the hospital in the past 90 days, recommend SLK; otherwise, recommend liver transplant alone. By representing the treatment regime as a decision list, the treatment regime is more understandable for both patients and providers when faced with choosing between SLK or liver transplant alone.

The genetic algorithm is well-suited for a wide range of optimization problems because it seeks to maximize a user-defined fitness function, meaning that users can tailor the algorithm for the quantity of their interest. For example, for the case of feature selection for prediction models, we used mean squared prediction error to evaluate how well the selected set of covariates was predicting outcomes. As we previously discussed in Section 2, the genetic algorithm typically requires inputs to be specified as a set of binary indicators (Goldberg and Holland, 1988). In the case of feature selection, each of these binary indicators represents whether or not a certain covariate was included in

building the prediction model. One major challenge of developing a genetic algorithm for treatment regimes defined as decision lists is the definition of these binary indicators. Genetic algorithms have been used for optimizing decision trees in previous literature. Some authors have proposed extending the binary indicators to include additional information such as the value of the attribute and the operator for joining the tree (Carvalho and Freitas, 2004; Liu and Fan, 2014). Other methods have included complex encoding to traverse the decision tree (Cha and Tappert, 2008). Because we aim only to optimize a decision list rather than a decision tree, we can simplify to avoid the complex encoding and decoding mechanism necessary for a decision tree. We propose defining these binary indicators to represent the inclusion or exclusion of clauses in the decision list. We propose a genetic algorithm-based optimization method for treatment regimes defined as decision rules and aim to take advantage of the genetic algorithm's ability to widely search across a problem space while narrowing in on the best-performing solution.

We were motivated by developing clinical guidelines for which patients who are awaiting liver transplant and who have decreased renal function should undergo simultaneous liver-kidney transplant (SLK) or liver alone transplant. Simultaneous liver kidney (SLK) transplantation has become increasingly common, particularly since the MELD score began to be used for liver allocation in 2002, increasing from 2.5% in 1994 to 10.3% in 2009 (Singal *and others*, 2013). However, for some patients, renal dysfunction may be reversible following improved health after liver transplantation alone. Therefore, identifying patients for whom SLK transplant would be most beneficial is crucial for the allocation of scarce kidneys. It is important in this context that the treatment strategies be simple and interpretable to improve physician and patient uptake.

Changes in the Organ Procurement Transplant Network (OPTN)'s allocation policy for SLK are as recent as 2017; limited analysis of the impact of these policies has been conducted. A recent literature review on SLK transplant criteria and outcomes

described the gaps in allocation criteria and indicated the need for further examination of how SLK allocation is conducted, particularly for those with acute kidney injury (Singal *and others*, 2019).

The rest of Section 4 is organized as follows. In Section 4.3.1 and 4.3.2, we present our notation, assumptions, and optimization method. In Section 4.3.3, we present our methods for the application to patients who may potentially benefit from a simultaneous liver and kidney transplant and in Section 4.4 we present the results. In Section 4.5, we discuss our conclusions and future extensions of this work.

## 4.3 Methods

### 4.3.1 Notation and Causal Assumptions

We focus our method and analysis assessing a treatment regime with a survival outcome; however, this method could be easily adapted to alternate outcomes as well. We observe data $(\boldsymbol{X}_i, A_i, V_i, \delta_i)$ for $i = 1, \ldots N$ individuals, where $\boldsymbol{X}$ are patient covariates that can be continuous or categorical and $A \in (0, 1)$ is the treatment assigned. We let $V_i = min(T_i, C_i)$ be the observed time until event prior to $\tau$, where $T_i$ is the time between baseline and the event of interest and $C_i$ is the time between baseline and loss to follow-up. We let $\delta_i = \mathbb{I}(T_i < C_i)$ be the observed indicator for the event of interest.

We define the propensity score $p(x) = P(A_i = 1|\boldsymbol{X}_i = x_i)$ and inverse probability of treatment weights as

$$\omega_i = \begin{cases} \frac{1}{p(x)}, & \text{if } a_i = 1 \\ \frac{1}{1-p(x)}, & \text{if } a_i = 0 \end{cases}$$

$\hat{p}(x)$ can be estimated by logistic regression, random forest, or alternate methods.

Let $V_i^a$ and $\delta_i^a$ denote the counterfactual survival outcomes that would be observed

for an individual under treatment $a$. To estimate the causal effect of a treatment regime on the survival probability, we make the following assumptions. We assume consistency, that is $V_i^a = V_i$ and $\delta_i^a = \delta_i$ for an individual with $A_i = a$. We also assume positivity, that is $0 < P(A_i = a) < 1$ for all values of $a$; we are assuming that there is a non-zero probability that a patient will receive the treatment and that there is a non-zero probability that a patient won't receive the treatment. Finally, we assume no unmeasured confounding, that is that the probability of treatment depends only on the observed data and not additionally on any potential outcomes.

### 4.3.2 Genetic algorithm for identifying optimal treatment regime

We define the treatment regime as a decision list of clauses selected by the genetic algorithm, where each clause is a logical condition that is true or false for each individual based on their covariates $x_i$. For example, one clause may be "if age is greater than 50". We define the decision list as the combination of each clause with an "or" statement, such that if any of the clauses are true for an individual, the regime indicates that they will receive the specified treatment.

To illustrate how we write covariates as clauses and then take those clauses as a decision list, we provide a simple example. Take the scenario in which there are only 2 covariates: $x_1$ which is continuous and $x_2$ which is categorical with 2 levels, $c_1$ and $c_2$. Of interest for the continuous variable $x_1$ are quartiles of its distribution in the data, $q_1$, $q_2$, and $q_3$. We then define the following clauses:

1. $x_1 \leq q_1$
2. $x_1 > q_1$ and $x_1 \leq q_2$
3. $x_1 > q_2$ and $x_1 \leq q_3$
4. $x_1 > q_3$

5. $x_2 = c_1$

6. $x_2 = c_2$

If, for example, our method yields clauses 1 and 5 as composing the best solution, then the decision list can be written as: if $x_1 \leq q_1$ or $x_2 = c_1$, then provide treatment, else, do not provide treatment.

Our method for identifying the optimal treatment regime, as described by a decision list, is displayed graphically in Figure 4.1 and is described by the following steps:

1. For features $\boldsymbol{X} \in \mathbb{R}^p$ describing patient covariates, identify $L$ potential splits to be used in the decision list clauses. For categorical attributes, define one clause per category or relevant groupings of categories. For numerical attributes, identify potential thresholds, recommended as quartiles of the distribution of the covariate within the training dataset, and define one clause per threshold.

2. Define a chromosome $C_{jk}$ composed of L genes $g_{jkl} \in \{0, 1\}$, where each entry $g_{jkl}$ indicates whether or not clause $l$ is selected for inclusion in the decision list. For $k = 1, \ldots, K$ chromosomes, initialize the first generation, $j = 1$, of $K$ chromosomes via random generation.

3. Evaluate each chromosome $C_{jk}$ under a fitness function $f(C_{jk})$ designed to evaluate the treatment regime given the selected clauses of the decision list indicated by the $g_{jkl}$ entries of the chromosome. Further description of the fitness function is described in the next section.

    i. For each individual $i$ in the training dataset, identify whether or not they are compliant to the rule under the clauses selected by $C_{jk}$.

    ii. Apply treatment regime indicated by the decision list to training dataset and calculate $f(C_{jk})$.

4. Generate the reproducing population of chromosomes for the following generation.

    i. Elitism: Automatically include the top-performing chromosomes as evaluated by the fitness function in the following generation. Include the top $K * p_E$ chromosomes, where $p_E$ is the proportion of elitism.

    ii. Selection: Select $K * (1 - p_E)$ high-performing chromosomes for inclusion in the reproducing population, using linear-rank selection. Exclude the $K * p_E$ chromosomes from Step 4ii from the selection process.

5. Apply genetic operators to the $K * (1 - p_E)$ chromosomes generated in step 4ii to form the next generation.

    i. Crossover: Randomly pair the reproducing population chromosomes to form $K*(1-p_E)/2$ pairs referred to as "parents". Apply uniform crossover between the two individuals with probability $p_C$.

    ii. Mutation: Randomly modify a small number of $g_{jkl}$ to induce genetic variation. Change the indicator of each 0/1 $g_{jkl}$ in each chromosome in the new population with probability $p_M$.

6. Repeat steps 3-5 for $j = 1, \ldots, N$ iterations, leading to a final generation of chromosomes. Select clauses for inclusion in the decision list per the chromosome with the best performance according to the fitness function $f$, denoted $C_{N*}$.

Elitism proportion for use in selection was set to $p_E = 0.05$, a commonly used value. We applied uniform crossover with a crossover probability $p_C = 0.5$ and mutation probability $p_M = 0.1$. A population size of $K = 200$ was chosen and the algorithm was repeated for $N = 200$ generations or until no improvement in the fitness function was made for over 20 generations.

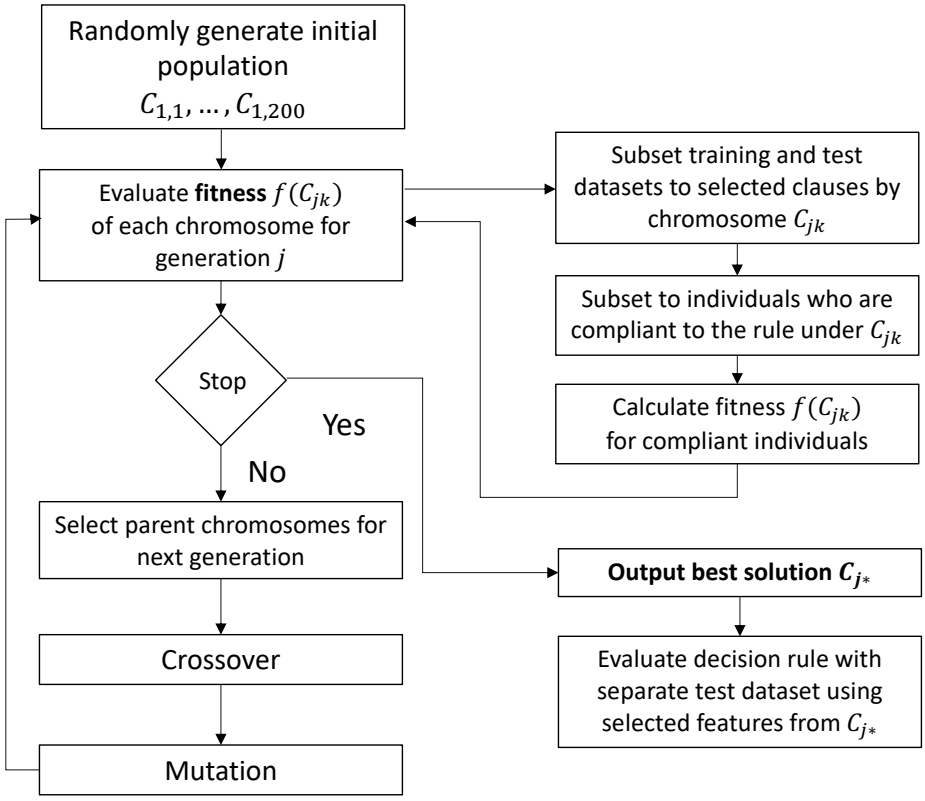


Figure 4.1: Flowchart detailing the method to identify treatment regimes using the genetic algorithm.

Most genetic algorithms begin the first generation with random generation using a uniform distribution. With the goal to identify a treatment regime that could be of clinical interest, we instead suggest the use of a Bernoulli distribution with $p =$ to generate each gene in the first generation, thus lowering the number of clauses included in the decision list.

### 4.3.3 Fitness function for survival outcomes with possible confounding

The genetic algorithm utilizes a fitness function at each iteration to evaluate the performance of each potential treatment regime. The choice of this fitness function can greatly impact the results of the algorithm's optimization. As introduced in Section 4.3.1, we develop our method initially for a survival outcome with observed time until event or loss to follow-up $V_i$ and observed event indicator $\delta_i$ for individual $i$.

For a survival outcome that requires adjustment for confounders, we recommend defining $f(C_{jk})$ as the adjusted restricted mean survival time (RMST) at time $\tau$. RMST is a useful summary measure for time-to-event data that represents the average event-free survival from time 0 to $\tau$ and is equal to the area under the survival curve up to time $\tau$. The adjusted RMST can be calculated by integrating a Kaplan-Meier curve adjusted with inverse probability weights (IPW); Conner *and others* (2019) present proof of concept and technical details for this method of calculating RMST when there is possible confounding. We describe here its use in our fitness function $f(C_{jk})$.

Recall that fitness function is evaluated for each potential chromosome $C_{jk}$, where $j$ denotes the generation and $k$ denotes the chromosome. Prior to the calculation of RMST, we subset the data to include only those individuals $i$ that are compliant with the treatment regime under the clauses selected by $C_{jk}$. Let $\gamma_i^{C_{jk}} \in \{0, 1\}$ be the indicator for compliance. For simplicity of notation, we denote $\gamma_i^{C_{jk}}$ as only $\gamma_i$ and let the following notation describe the calculation of the fitness function for one chromosome.

For D distinct times at which events occur $t_1, \ldots, t_D$, let $m = 1, \ldots, D$. Let $d_m = \sum_{i:V_i=t_m} \delta_i \gamma_i$ be the number of events among compliant individuals and $Y_m = \sum_{i:V_i \geq t_m} \gamma_i$ be the number of compliant individuals at risk at time $t_m$. The standard Kaplan-Meier estimator of survival at time $t$ is then given by $\hat{S}(t) = \prod_{t_m \leq t} \left[1 - \frac{d_m}{Y_m}\right]$.

To adjust for possible confounding, we utilize IPW; we defined the inverse probability of treatment weights $\omega_i$ in Section 4.3.1. We apply the weights to the survival distribution to get $\tilde{d}_m = \sum_{i:V_i=t_m} \omega_i \delta_i \gamma_i$ and $\tilde{Y}_m = \sum_{i:V_i \geq t_m} \omega_i \gamma_i$. These then yield the adjusted Kaplan-Meier estimate $\hat{S}_{adj}(t) = \prod_{t_m \leq t} \left[1 - \frac{\tilde{d}_m}{\tilde{Y}_m}\right]$. Finally, the IPW-adjusted RMST at time $\tau$ is calculated as $\hat{\mu}_{adj}(\tau) = \int_0^\tau \hat{S}_{adj}(t)dt$.

### 4.3.4 Application to simultaneous liver kidney transplant

As first introduced in Section 1.2.3, we aim to identify a decision rule for whether or not a given patient listed for liver transplant should undergo liver transplant alone or SLK transplant. Data for this analysis were collected from the United Network for Organ Sharing national registry. Transplants were included if they occurred between February 1st, 2002 and June 1st, 2017; this time frame was selected to include patients for whom the MELD score was used for allocation and whose 3 year follow-up data would be available. These transplants also occurred prior to the implementation of the OPTN's updated allocation criteria for SLK transplant. Transplants in our analysis were restricted to those for patients aged 18 or older at the time of transplant. Patients who had received a previous transplant or received organs outside of the liver or kidney were excluded. In order to ensure that it would be reasonable for the patient to receive both a liver and kidney transplant simultaneously, we subset to those with mediocre kidney function, defined in our analysis as having an eGFR, derived from the CKD-EPI formula, of less than 60 mL/min/1.73 $m^2$ or being on dialysis at the time of transplant. Estimated glomerular filtration rate (eGFR) is a measure of kidney function based on the patient's creatinine, age, race, and gender. Organ procurement organizations (OPO) with fewer than 5% of their liver transplants consisting of simultaneous liver and kidney transplant during our analysis period were excluded in order to avoid confounding; 3 OPOs were excluded.

33,658 liver transplants meet our inclusion and exclusion criteria for analysis. Of these, 5,419 (16.1%) are simultaneous liver and kidney transplants. 21 covariates regarding transplant recipient demographics and health status are included. Descriptive characteristics are presented in Table 4.1.

We consider the joint outcome of liver graft failure and mortality at 3 years post-transplant. 7,696 patients (22.9%) included in our analysis observed this event. An additional 3,171 patients (9.4%) were lost to follow-up during the 3-year follow-up period. The outcome was weighted for use in analysis by inverse probability of treatment, where treatment is SLK; weights were built for the probability of treatment by random forest using the patient characteristics described in Table 4.1. We used a fitness function of IPW-adjusted restricted mean survival time (RMST) as described in Section 4.3.3 for $\tau$ of 3 years.

Due to the scarcity of organs available for transplant, we conducted additional analysis by restricting the total proportion of SLK recommended by the regime. We also conduct a secondary analysis where we add 50 additional days of survival to those who did not receive SLK, in order to identify treatment regimes that provide at least 50 days of benefit for those assigned to SLK.

## 4.4 Results

Under no restriction to the population level of transplant, the treatment regime identified by the genetic algorithm recommended 98% of individuals in our target population to receive SLK rather than liver transplant alone. 17 clauses were included in this decision list, as described in Table 4.2; recall that the clauses listed are joined by "or" connectors, such that a patient meeting any of the listed criteria would be recommended for SLK. These results are not surprising; because we limited our population of interest

in the analysis to patients already on dialysis or with mediocre kidney function defined by eGFR, we expect to see that most of these patients would receive some survival benefit from SLK.

Restricting the population level of SLK to 50% led to 3 clauses selected for inclusion in the decision list: patients were recommended to receive SLK if they had been on dialysis for greater than 2 days or if their last albumin value was under 2.5 or if their liver diagnosis did not fall into one of the broad categories defined by the OPTN as reasons for liver transplant. Restricting the population level of SLK to 25% led to 2 clauses selected for inclusion in the decision list and a decision rule of any patient who is not white being recommended to receive SLK. We found that the 3-year restricted mean survival time (RMST) was highest when no restrictions were made to the level of SLK permitted in the population and decreased as the level of restriction increased.

In Table 4.3, we present the results of our method when we added 50 additional days of survival to those who did not receive SLK, in order to identify treatment regimes that offer at least that much benefit for SLK. We note that the 3-year RMST displayed in Tables 4.2 and Table 4.3 are not directly comparable due to our modification. We found similar results as in our unmodified analysis: both 3-year RMST and the number of clauses included in the treatment regime increased as the level of restriction on SLK in the population decreased, with one exception: the treatment regime under maximum 75% SLK in our secondary analysis was as good as the unrestricted version. In addition, the range of RMST values was smaller in our secondary analysis than in the unmodified version.

| **Categorical Patient Characteristics** | N (%) |
|---|---|
| Gender | |
| Female | 14,552 (39.2) |
| Male | 22,525 (60.8) |

| | |
|---|---|
| Race | |
| Black | 3,545 (9.6) |
| Other | 1,747 (4.7) |
| White | 31,785 (85.7) |
| Blood type | |
| A | 13,754 (37.1) |
| AB | 1,641 (4.4) |
| B | 4,821 (13.0) |
| O | 16,802 (45.3) |
| Other | 59 (0.2) |
| Liver diagnosis category | |
| Acute hepatic necrosis | 1,945 (5.2) |
| Metabolic diseases | 846 (2.3) |
| Malignant neoplasms | 1,779 (4.8) |
| Cholestatic liver disease/cirrhosis | 2,582 (7.0) |
| Non-cholestatic liver disease | 26,752 (72.2) |
| Biliary atresia | 3,096 (8.4) |
| Other | 77 (0.2) |
| Last ascites | |
| Absent | 4942 (13.3) |
| Slight | 15,991 (43.1) |
| Moderate | 16,144 (43.5) |
| Last encephalopathy | |
| None | 8,755 (23.6) |
| 1–2 | 21,332 (57.5) |
| 3–4 | 6,990 (18.9) |
| On dialysis the week prior | |
| Yes | 10,294 (27.8) |
| No | 26,783 (72.2) |
| On life support | |
| Yes | 5,628 (15.2) |
| No | 31,449 (84.8) |
| In the hospital in the past 90 days | |
| Yes | 14,018 (37.8) |
| No | 8,035 (21.7) |
| Unknown | 15,024 (40.5) |
| Diabetes diagnosis | |

| | |
|---|---|
| Yes | 1,366 (3.7) |
| No | 35,711 (96.3) |
| Any previous Malignancy | |
| Yes | 3,600 (9.7) |
| No | 32,275 (87.0) |
| Unknown | 1,202 (3.2) |
| Drug treated systemic hypertension | |
| Yes | 7,342 (19.8) |
| No | 18,706 (50.5) |
| Unknown | 11,029 (29.7) |
| Working for income | |
| Yes | 2,625 (7.1) |
| No | 29,749 (80.2) |
| Unknown | 4,703 (12.7) |

| **Continuous Patient Characteristics** | Mean (sd) |
|---|---|
| Age (years) | 55.1 (10.04) |
| Body Mass Index (kg/$m^2$) | 28.65 (6.04) |
| Height (cm) | 171.20 (10.31) |
| Weight (kg) | 84.16 (20.09) |
| Time on dialysis (days) | 13.99 (85.92) |
| Last serum creatine (mg/dL) | 2.47 (1.59) |
| Last bilirubin (mg/dL) | 12.47 (13.75) |
| Last albumin (g/dL) | 3.02 (0.77) |
| Last INR (International Normalized Ratio) | 2.12 (1.53) |
| eGFR (mL/min/1.73 $m^2$) | 36.28 (20.15) |

Table 4.1: The mean and standard deviation (if continuous) or frequency and proportion (if categorical) of each covariate are presented for the 37,077 patients used in the analysis, where "last" refers to the most recent value of the variable at the time of transplant.

| Maximum Percent SLK | 3-year RMST (95% CI) | Selected Treatment Regime |
|---|---|---|
| 100 | 935.8 (931.5, 940.2) | Height $\leq$ 178 and $>$ 164, weight $\leq$ 82 and $>$ 70, last ascites moderate, Last INR $>$ 1.4 and $\leq$ 2.8, last albumin $\leq$ 3.5 and $>$ 3, cholestatic liver disease/cirrhosis or other, has previous malignancy, has drug treated hypertension, working for income or unknown, not in the hospital in the past 90 days |
| 90 | 934.0 (929.9, 938.1) | On life support, last INR $\leq$ 1.4, last albumin $\leq$ 2.5, last serum creatinine $>$ 1.9, has diabetes, AB blood type, cholestatic liver disease/cirrhosis |
| 75 | 926.6 (922.4, 930.7) | Female, Black, last ascites absent, last serum creatinine $>$ 3, last albumin $\leq$ 2.5 |
| 50 | 914.1 (910.1, 918.1) | Last albumin $\leq$ 2.5, No liver diagnosis category, time on dialysis $>$ 2 days |
| 25 | 879.8 (875.7, 883.9) | Black or Other race |
| 10 | 876.9 (872.9, 881.0) | Black |

Table 4.2: Selected treatment regimes for assigning SLK and 3-year RMST (with 95% confidence interval) via our genetic algorithm method under varying restrictions of SLK.

## 4.5 Discussion

We made multiple simplifying assumptions in our definition of the treatment regime as a decision list. Primarily, we focused only on "or" statements to join the clauses of

| Maximum Percent SLK | 3-year RMST (95% CI) | Selected Treatment Regime |
|---|---|---|
| 100 | 972.3 (967.7, 976.8) | Height $\leq$ 164 cm, weight $>$ 83 kg, on life support, on dialysis the prior week, last bilirubin $>$ 19.2, last serum creatinine $>$ 3, last INR $\leq$ 1.8 and $>$ 2.4, eGFR $\leq$ 35.3 and $>$ 20.3 |
| 90 | 971.9 (967.5, 976.3) | Weight $>$ 70 and $\leq$ 82, age $>$ 50, on dialysis the prior week, last ascites absent, last bilirubin $\leq$ 2.5, last albumin $\leq$ 2.5, not in the hospital in the previous 90 days, working for income, has drug treated hypertension |
| 75 | 972.5 (968.2, 976.8) | Female, Black, last albumin $\leq$ 2.5, last serum creatinine $>$ 3, last ascites absent |
| 50 | 959.9 (955.9, 963.9) | Last bilirubin $\leq$ 2.5, has diabetes, working for income, eGFR $\leq$ 20.7 |
| 25 | 943.2 (939.2, 947.2) | Black or Other race |
| 10 | 941.0 (937.0, 945.0) | Black |

Table 4.3: Selected treatment regimes for assigning SLK and 3-year RMST (with 95% confidence interval) via our genetic algorithm method under varying restrictions of SLK, when 50 additional days of survival are added for those who did not receive SLK.

the decision list. This limits our potentially selected treatment regimes by excluding "interaction" clauses that might be relevant. For example, a treatment regime that suggests patients with high serum creatinine values should receive SLK, but only if they also are older than 50, would not be captured by our method. The use of "or" clauses combined with a lack of penalization on the number of clauses selected by our genetic algorithm might also lead to the inclusion of unnecessary clauses that overlap

heavily with another selected clause. In addition, we limited analysis to two potential treatments; in alternate contexts, more than one treatment may be of interest.

Because of our interest in a survival outcome from observational data, we used an adjusted restricted mean survival estimate as the fitness function. In a clinical trials context with uncensored outcomes, sample means could be used as the fitness function. Future work includes further investigation of our method's performance under alternate choices of fitness function.

As is the case with most questions in organ transplant, we must also consider the limiting factor of available kidneys for transplant. Our results for treatment regimes for SLK reinforced the challenges in allocating kidneys to liver transplant patients. Without any restriction on the level of SLK allowed in the population, our results indicate that nearly every patient with existing renal dysfunction would benefit from a kidney transplant in addition to their liver transplant. While we focus here on identifying patients with liver transplants for the additional kidney transplant, we must note that any kidneys allocated to these patients are not being allocated to patients without liver disease who are in need of kidneys for transplant. This analysis focused on SLK transplant among patients listed for liver transplant; patients listed for kidney transplant can also receive SLK. Although we focused here on identifying treatment regimes for those patients listed for liver transplant, any regimes implemented on this portion of the population impact the entire allocation of both SLK and kidney alone transplants; further examination of the population impact via the allocation process is warranted.

# References


Alagoz, Oguzhan, Maillart, Lisa M., Schaefer, Andrew J. and Roberts, Mark S. (2007). Determining the acceptance of cadaveric livers using an implicit model of the waiting list. *Operations Research* **55**(1), 24–36.

Allen, Genevera I. (2013). Automatic feature selection via weighted kernels and regularization. *Journal of Computational and Graphical Statistics* **22**(2), 284–299.

Blickle, Tobias and Thiele, Lothar. (1995). A Comparison of Selection Schemes used in Genetic Algorithms. *Technical Report* 11, Computer Engineering and Communication Networks Lab TIK.

Boatman, Jeffrey A. and Vock, David M. (2018). Estimating the causal effect of treatment regimes for organ transplantation. *Biometrics* **74**(4), 1407–1416.

Carvalho, Deborah R and Freitas, Alex A. (2004). A hybrid decision tree/genetic algorithm method for data mining. *Information Sciences* **163**(1-3), 13–35.

Cha, Sung Hyuk and Tappert, Charles. (2008). Constructing binary decision trees using genetic algorithms. In: *Proceedings of the 2008 International Conference on Genetic and Evolutionary Methods, GEM 2008*. pp. 49–54.

Conner, Sarah C., Sullivan, Lisa M., Benjamin, Emelia J., LaValley, Michael P., Galea, Sandro and Trinquart, Ludovic. (2019). Adjusted restricted mean survival times in observational studies. *Statistics in Medicine* **38**(20), 3832–3860.

Cortes, Corinna and Vapnik, Vladimir. (1995). Support-Vector Networks. *Machine Learning* **20**(973), 273–297.

Dasgupta, Sayan, Goldberg, Yair and Kosorok, Michael R. (2019). Feature elimination in kernel machines in moderately high dimensions. *Annals of Statistics* **47**(1), 497–526.

Dasgupta, Sayan and Huang, Ying. (2020). Selecting biomarkers for building optimal treatment selection rules by using kernel machines. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **69**(1), 69–88.

Drucker, Harris, Burges, Chris J C, Kaufman, Linda, Smola, Alex and Vapnik, Vladimir. (1997). Support Vector Regression Machines. In: *Advances in Neural Information Processing Systems*. pp. 155–161.

Fu, Haoda, Zhou, Jin and Faries, Douglas E. (2016). Estimating optimal treatment regimes via subgroup identification in randomized control trials and observational studies. *Statistics in Medicine* **35**(19), 3285–3302.

Goldberg, David E. and Holland, John H. (1988). Genetic Algorithms and Machine Learning. *Machine Learning* **3**(2), 95–99.

Han, Song, Pool, Jeff, Tran, John and Dally, William. (2015). Learning both weights and connections for efficient neural network. In: *Advances in Neural Information Processing Systems*, Volume 28. Curran Associates, Inc.

Hernán, Miguel Ángel, Brumback, Babette and Robins, James M. (2000). Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology* **11**(5), 561–570.

Horn, Jeffrey, Nafpliotis, Nicholas and Goldberg, David E. (1994). A niched Pareto genetic algorithm for multiobjective optimization. In: *Proceedings of the First IEEE Conference on Evolutionary Computation, IEEE World Congress on Computational Intelligence*, Volume 1. Piscataway, New Jersey: IEEE Service Center. pp. 82–87.

Hou, Edwin S.H., Ansari, Nirwan and Ren, Hong. (1994). A Genetic Algorithm for Multiprocessor Scheduling. *IEEE Transactions on Parallel and Distributed Systems* **5**(2), 113–120.

Ibrahim, Hassan N., Foley, Robert, Tan, LiPing, Rogers, Tyson, Bailey, Robert F., Guo, Hongfei, Gross, Cynthia R. and Matas, Arthur J. (2009). Long-Term Consequences of Kidney Donation. *New England Journal of Medicine* **360**(5), 459–469.

Joffe, Marshall M., Yang, Wei Peter and Feldman, Harold. (2012). G-Estimation and Artificial Censoring: Problems, Challenges, and Applications. *Biometrics* **68**(1), 275–286.

Kharrat, Ahmed, Gasmi, Karim and Messaoud, Mohamed B E N. (2010). A Hybrid Approach for Automatic Classification of Brain MRI Using Genetic Algorithm and Support Vector Machine. *Leonardo Journal of Sciences* **9**(17), 71–82.

Kim, Sang Phil, Gupta, Diwakar, Israni, Ajay K. and Kasiske, Bertram L. (2015). Accept/decline decision module for the liver simulated allocation model. *Health Care Management Science* **18**(1), 35–57.

Laber, E. B. and Zhao, Y. Q. (2015). Tree-based methods for individualized treatment regimes. *Biometrika* **102**(3), 501–514.

Leardi, R., Boggia, R. and Terrile, M. (1992). Genetic algorithms as a strategy for feature selection. *Journal of Chemometrics* **6**(5), 267–281.

Li, Li, Jiang, Wei, Li, Xia, Moser, Kathy L., Guo, Zheng, Du, Lei, Wang, Qiuju, Topol, Eric J., Wang, Qing and Rao, Shaoqi. (2005). A robust hybrid between genetic algorithm and support vector machine for extracting an optimal feature gene subset. *Genomics* **85**(1), 16–23.

Liu, Dong Sheng and Fan, Shu Jiang. (2014). A modified decision tree algorithm based on genetic algorithm for mobile user classification problem. *The Scientific World Journal* **2014**.

Liu, Shuangyin, Tai, Haijiang, Ding, Qisheng, Li, Daoliang, Xu, Longqin and Wei, Yaoguang. (2013). A hybrid approach of support vector regression with genetic algorithm optimization for aquaculture water quality prediction. *Mathematical and Computer Modelling* **58**(3-4), 458–465.

Maldonado, Sebastian and Weber, Richard. (2010). Feature selection for support vector regression via Kernel penalization. In: *Proceedings of the International Joint Conference on Neural Networks*. IEEE. pp. 1–7.

Maldonado, Sebastián, Weber, Richard and Basak, Jayanta. (2011). Simultaneous feature selection and classification using kernel-penalized support vector machines. *Information Sciences* **181**(1), 115–128.

Maulik, Ujjwal and Bandyopadhyay, Sanghamitra. (2000). Genetic algorithm-based clustering technique. *Pattern Recognition* **33**(9), 1455–1465.

Moodie, Erica E.M., Dean, Nema and Sun, Yue Ru. (2014). Q-Learning: Flexible Learning About Useful Utilities. *Statistics in Biosciences* **6**(2), 223–243.

Moodie, Erica E. M., Richardson, Thomas S. and Stephens, David A. (2007). Demystifying Optimal Dynamic Treatment Regimes. *Biometrics* **63**(2), 447–455.

O'Leary, J. G., Levitsky, J., Wong, F., Nadim, M. K., Charlton, M. and Kim, W. R. (2016). Protecting the Kidney in Liver Transplant Candidates: Practice-Based Recommendations From the American Society of Transplantation Liver and Intestine Community of Practice.

Osowski, Stanislaw, Siroi, Robert, Markiewicz, Tomasz and Siwek, Krzysztof. (2009). Application of Support Vector Machine and Genetic Algorithm for Improved Blood Cell Recognition. *IEEE Transactions on Intrumentation and Measurement* **58**(7), 2159–2168.

Peng, Sihua, Xu, Qianghua, Ling, Xuefeng Bruce, Peng, Xiaoning, Du, Wei and Chen, Liangbiao. (2003). Molecular classification of cancer types from microarray data using the combination of genetic algorithms and support vector machines. *FEBS Letters* **555**(2), 358–362.

Robins, James M., Blevins, Donald, Ritter, Grant and Wulfsoh, Michael. (1992). G-estimation of the effect of prophylaxis therapy for pneumocystis carinii pneumonia on the survival of AIDS patients. *Epidemiology* **3**(4), 319–336.

Saravanan, P and Sailakshmi, P. (2015). Missing value imputation using fuzzy possibilistic c means optimized with support vector regression and genetic algorithm. *Journal of Theoretical and Applied Information Technology* **72**(1), 34–39.

Schaubel, Douglas E., Wolfe, Robert A. and Port, Friedrich K. (2006).

A Sequential Stratification Method for Estimating the Effect of a Time-Dependent Experimental Treatment in Observational Studies. *Biometrics* **62**(3), 910–917.

Shortreed, Susan M., Laber, Eric, Lizotte, Daniel J., Stroup, T. Scott, Pineau, Joelle and Murphy, Susan A. (2011). Informing sequential clinical decision-making through reinforcement learning: An empirical study. *Machine Learning* **84**(1-2), 109–136.

Singal, Ashwani K., Guturu, Praveen, Hmoud, Bashar, Kuo, Yong Fang, Salameh, Habeeb and Wiesner, Russell H. (2013). Evolving frequency and outcomes of liver transplantation based on etiology of liver disease. *Transplantation* **95**(5), 755–760.

Singal, Ashwani K., Ong, Song, Satapathy, Sanjaya K., Kamath, Patrick S. and Wiesner, Russel H. (2019). Simultaneous liver kidney transplantation.

Smola, Alex J and Scholkopf, Bernhard. (2004). A tutorial on support vector regression. *Statistics and Computing* **14**, 199–222.

Su, Xuanming, Zenios, Stefanos A. and Chertow, Glenn M. (2004). Incorporating recipient choice in kidney transplantation. *Journal of the American Society of Nephrology* **15**(6), 1656–1663.

Syswerda, Gilbert. (1989). Uniform crossover in genetic algorithms. In: *Proceedings of the Third International Conference on Genetic Algorithms*. Morgan Kaufmann Publishers. pp. 2–9.

Tibshirani, Robert. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **58**(1), 267–288.

Westreich, Daniel, Cole, Stephen R., Young, Jessica G., Palella, Frank, Tien, Phyllis C., Kingsley, Lawrence, Gange, Stephen J and Hernán, Miguel A. (2012). The parametric g-formula to estimate the effect of highly active antiretroviral therapy on incident AIDS or death. *Statistics in Medicine* **31**(18), 2000–2009.

Wu, Chih Hung, Tzeng, Gwo Hshiung and Lin, Rong Ho. (2009). A Novel hybrid genetic algorithm for kernel function and parameter optimization in support vector regression. *Expert Systems with Applications* **36**(3 PART 1), 4725–4735.

Yang, J. and Honavar, V. (1998). Feature subset selection using a genetic algorithm. *IEEE Intelligent Systems* **13**(2), 44–49.

Yang, Jian Bo and Ong, Chong Jin. (2011). Feature selection using probabilistic prediction of support vector regression. *IEEE Transactions on Neural Networks* **22**(6), 954–962.

Zhang, Baqun, Tsiatis, Anastasios A., Laber, Eric B. and Davidian, Marie. (2012). A Robust Method for Estimating Optimal Treatment Regimes. *Biometrics* **68**(4), 1010–1018.

Zhang, Baqun and Zhang, Min. (2018). C-learning: A new classification framework to estimate optimal dynamic treatment regimes. *Biometrics* **74**(3), 891–899.

Zhang, Yichi, Laber, Eric B., Tsiatis, Anastasios and Davidian, Marie. (2015). Using decision lists to construct interpretable and parsimonious treatment regimes. *Biometrics* **71**(4), 895–904.

Zhao, Yingqi, Zeng, Donglin, Rush, A. John and Kosorok, Michael R. (2012). Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association* **107**(499), 1106–1118.

Zou, Hui. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101**(476), 1418–1429.

# Appendix A

# Appendix

## A.1 Simulation methods details

We present additional technical details of the simulation study presented in Section 2.4 of the full text. Data generation schemes and results were presented in Section 2.4; here, we present the specific settings used for the simulation results presented in the text.

In scenario 1, data was generated to produce a quadratic relationship between outcome and covariates. The amount of the relationship that was quadratic vs. linear was varied. Values of $\beta$ were chosen such that approximately 50% of the variation in the outcome was attributable to the relationship with the covariates. The first set of simulations was run with $p = 50$ and $\pi = 0.1$, indicating 50 covariates with 5 of them associated with the outcome. The following sets of $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$, where $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are the coefficients for the quadratic and linear terms respectively, were used: $(0.32, 0), (0.26, 0.26), (0.18, 0.365), (0, 0.45)$. The second set of simulations was run with $p = 300$, $\pi = 0.05$, and the following values of $\boldsymbol{\beta}$: $(0.18, 0), (0.15, 0.15), (0.105, 0.21), (0, 0.26)$. Results of scenario 1 are presented in Figures 2 and 3 of the full text and Section 2 of the Appendix.

In scenario 2, data was generated such that interaction terms were associated with the outcome. The following settings were used for the results presented in Table 1 of the full text: $p = 50, \pi = 0.1, \boldsymbol{\beta} = (0.22, 0.32)$.

In scenario 3, covariates were correlated multivariate normal instead of uncorrelated as in scenarios 1 and 2. We set $p = 50$, $\pi = 0.1$, and $\rho = 0.5$. Two block covariate settings were designed, one in which main effect terms were correlated with terms unassociated with the outcome where $\beta = 0.45$ for the main effect covariates, and one in which the main effect terms were correlated with each other where $\beta = 0.32$ for the main effect covariates. Results of these settings are presented in Tables 2 and 3 of the main text respectively.

## A.2 Additional simulation results

Complete results of simulations for scenario 1, in which the true relationship between covariates and outcomes is an underlying quadratic relationship, are presented. The relative weight of the quadratic association vs. the linear association is varied from an entirely quadratic association to an entirely linear one; the quadratic association is considered underlying and is not supplied to the model as a predictor. The number of covariates $p$ considered are 50 and 300; tables 1-4 below correspond to $p = 50$ with $\pi = 0.1$ and tables 5-8 correspond to $p = 300$ with $\pi = 0.05$, where $\pi$ is the proportion of covariates truly related to the outcome. Specific values of the coefficients $\boldsymbol{\beta}$ are presented in Section 1 of the Appendix, above. Results are described in the text in Section 2.4.2 and correspond to Figures 2 and 3 of the full text.

The column described as "Avg % Variation Unexplained by Pred" displays the average across simulation iterations of the proportion of true variation in $\mathbf{Y}$ due to $\mathbf{X}$ that is left unexplained by the prediction. For example, in the case where the total

variance in $\mathbf{Y}$ is 2, where half is due to the error term and half is due to the covariates $\mathbf{X}$, a MSPE of 1.05 would be reported as 5% in this column. Of the explainable variance 1, the model misses 5%. In some circumstances, this column may be listed as greater than 100%, indicating that the MSPE is higher than the total amount of variation in the outcome.

| Feature Selection | Model Fit | % Selection Relevant (1+) | Avg # Relevant Var | % Other Var (1+) | Avg # Additional Var | Avg % Variation Unexplained by Pred. |
|---|---|---|---|---|---|---|
| GA: MSE | LR | 87.8 | 1.72 | 100 | 15.68 | 106.29 |
| GA: BIC | LR | 17.6 | 0.18 | 82.4 | 0.83 | 103.49 |
| GA: MSE | SVR | 100 | 5 | 44.8 | 0.70 | 7.03 |
| GA: BIC | SVR | 100 | 5 | 14.6 | 0.18 | 6.10 |
| RFE | SVR | 100 | 4.93 | 100 | 3.08 | 22.33 |
| LASSO | LR | 0 | 0 | 0 | 0 | 103.31 |
| None | RF | 100 | 5 | 100 | 45 | 23.49 |
| None | SVR | 100 | 5 | 100 | 45 | 74.83 |

Table A1: Results of simulation 1 setting with 50 total covariates X, 5 of them truly associated with the outcome Y with entirely quadratic relationship, where the underlying quadratic effects not supplied to the model for selection or prediction.

| Feature Selection | Model Fit | % Selection Relevant (1+) | Avg # Relevant Var | % Other Var (1+) | Avg # Additional Var | Avg % Variation Unexplained by Pred. |
|---|---|---|---|---|---|---|
| GA: MSE | LR | 100 | 4.98 | 100 | 15.68 | 72.31 |
| GA: BIC | LR | 100 | 4.97 | 6.4 | 0.07 | 70.24 |
| GA: MSE | SVR | 100 | 5.0 | 35.4 | 0.62 | 8.76 |
| GA: BIC | SVR | 100 | 5.0 | 14.0 | 0.17 | 8.02 |
| RFE | SVR | 100 | 5.0 | 100 | 3.0 | 19.38 |
| LASSO | LR | 100 | 5.0 | 2.6 | 0.03 | 69.97 |
| None | RF | 100 | 5.0 | 100 | 45 | 19.43 |
| None | SVR | 100 | 5.0 | 100 | 45 | 56.82 |

Table A2: Results of simulation 1 setting with 50 total covariates X, 5 of them truly associated with the outcome Y with mostly quadratic relationship and smaller linear relationship, where the underlying quadratic effects not supplied to the model for selection or prediction.

| Feature Selection | Model Fit | % Selection Relevant (1+) | Avg # Relevant Var | % Other Var (1+) | Avg # Additional Var | Avg % Variation Unexplained by Pred. |
|---|---|---|---|---|---|---|
| GA: MSE | LR | 100 | 5 | 100 | 15.96 | 36.78 |
| GA: BIC | LR | 100 | 5 | 6.8 | 0.07 | 35.01 |
| GA: MSE | SVR | 100 | 5 | 42.8 | 0.79 | 10.17 |
| GA: BIC | SVR | 100 | 5 | 5.6 | 0.06 | 9.46 |
| RFE | SVR | 100 | 5 | 100 | 3.12 | 18.27 |
| LASSO | LR | 100 | 5 | 2.6 | 0.03 | 34.97 |
| None | RF | 100 | 5 | 100 | 45 | 15.28 |
| None | SVR | 100 | 5 | 100 | 45 | 34.46 |

Table A3: Results of simulation 1 setting with 50 total covariates X, 5 of them truly associated with the outcome Y with mostly linear relationship and smaller quadratic relationship, where the underlying quadratic effects not supplied to the model for selection or prediction.

| Feature Selection | Model Fit | % Selection Relevant (1+) | Avg # Relevant Var | % Other Var (1+) | Avg # Additional Var | Avg % Variation Unexplained by Pred. |
|---|---|---|---|---|---|---|
| GA: MSE | LR | 100 | 5 | 100 | 15.74 | 2.97 |
| GA: BIC | LR | 100 | 5 | 6.0 | 0.07 | 1.63 |
| GA: MSE | SVR | 100 | 5 | 100 | 13.14 | 3.99 |
| GA: BIC | SVR | 100 | 5 | 22.0 | 0.24 | 2.55 |
| RFE | SVR | 100 | 5 | 100 | 3 | 16.55 |
| LASSO | LR | 100 | 5 | 3.0 | 0.03 | 1.63 |
| None | RF | 100 | 5 | 100 | 45 | 16.08 |
| None | SVR | 100 | 5 | 100 | 45 | 7.54 |

Table A4: Results of simulation 1 setting with 50 total covariates X, 5 of them truly associated with the outcome Y with entirely linear relationship.

| Feature Selection | Model Fit | % Selection Relevant (1+) | Avg # Relevant Var | % Other Var (1+) | Avg # Additional Var | Avg % Variation Unexplained by Pred. |
|---|---|---|---|---|---|---|
| GA: MSE | LR | 99.4 | 4.5 | 100 | 83.97 | 147.2 |
| GA: BIC | LR | 8.0 | 0.08 | 92.5 | 0.96 | 129.1 |
| GA: MSE | SVR | 100 | 15 | 71.5 | 2.93 | 18.9 |
| GA: BIC | SVR | 100 | 15 | 48.0 | 0.83 | 15.1 |
| RFE | SVR | 0.96 | 4.10 | 100 | 40.94 | 103.0 |
| LASSO | LR | 0 | 0 | 0 | 0 | 128.9 |
| None | RF | 100 | 15 | 100 | 285 | 72.2 |
| None | SVR | 100 | 15 | 100 | 285 | 126.9 |

Table A5: Results of simulation 1 setting with 300 total covariates X, 15 of them truly associated with the outcome Y with entirely quadratic relationship, where the underlying quadratic effects not supplied to the model for selection or prediction.

| Feature Selection | Model Fit | % Selection Relevant (1+) | Avg # Relevant Var | % Other Var (1+) | Avg # Additional Var | Avg % Variation Unexplained by Pred. |
|---|---|---|---|---|---|---|
| GA: MSE | LR | 100 | 22.66 | 100 | 79.18 | 119.9 |
| GA: BIC | LR | 99.7 | 4.71 | 26.1 | 0.31 | 110.3 |
| GA: MSE | SVR | 100 | 14.99 | 92.0 | 5.62 | 20.3 |
| GA: BIC | SVR | 100 | 14.5 | 70.5 | 1.24 | 15.3 |
| RFE | SVR | 100 | 12.36 | 100 | 32.64 | 61.4 |
| LASSO | LR | 7.2 | 0.35 | 03.9 | 0.15 | 113.2 |
| None | RF | 100 | 15 | 100 | 285 | 62.6 |
| None | SVR | 100 | 15 | 100 | 285 | 108.7 |

Table A6: Results of simulation 1 setting with 300 total covariates X, 15 of them truly associated with the outcome Y with mostly quadratic relationship and smaller linear relationship, where the underlying quadratic effects not supplied to the model for selection or prediction.

| Feature Selection | Model Fit | % Selection Relevant (1+) | Avg # Relevant Var | % Other Var (1+) | Avg # Additional Var | Avg % Variation Unexplained by Pred. |
|---|---|---|---|---|---|---|
| GA: MSE | LR | 100 | 14.83 | 100 | 72.93 | 37.63 |
| GA: BIC | LR | 100 | 14.75 | 57.7 | 0.84 | 29.62 |
| GA: MSE | SVR | 100 | 14.98 | 100 | 33.61 | 28.22 |
| GA: BIC | SVR | 100 | 14.98 | 85.0 | 1.70 | 20.95 |
| RFE | SVR | 100 | 14.68 | 100 | 30.3 | 33.43 |
| LASSO | LR | 100 | 14.96 | 95.4 | 6.03 | 29.49 |
| None | RF | 100 | 15 | 100 | 285 | 37.87 |
| None | SVR | 100 | 15 | 100 | 285 | 50.60 |

Table A7: Results of simulation 1 setting with 300 total covariates X, 15 of them truly associated with the outcome Y with mostly linear relationship and smaller quadratic relationship, where the underlying quadratic effects not supplied to the model for selection or prediction.

| Feature Selection | Model Fit | % Selection Relevant (1+) | Avg # Relevant Var | % Other Var (1+) | Avg # Additional Var | Avg % Variation Unexplained by Pred. |
|---|---|---|---|---|---|---|
| GA: MSE | LR | 100 | 14.97 | 100 | 71.84 | 9.79 |
| GA: BIC | LR | 100 | 14.96 | 68.4 | 1.18 | 3.46 |
| GA: MSE | SVR | 100 | 14.99 | 100 | 53.94 | 10.64 |
| GA: BIC | SVR | 100 | 14.98 | 92.0 | 2.19 | 4.62 |
| RFE | SVR | 100 | 15 | 100 | 30.0 | 16.88 |
| LASSO | LR | 100 | 15 | 96.7 | 6.02 | 3.74 |
| None | RF | 100 | 15 | 100 | 285 | 33.07 |
| None | SVR | 100 | 15 | 100 | 285 | 23.84 |

Table A8: Results of simulation 1 setting with 300 total covariates X, 15 of them truly associated with the outcome Y with entirely linear relationship.

## A.3 Lower signal-to-noise ratio simulations

As mentioned in Section 1 of the Appendix, for the simulation results presented in the text, approximately 50% of the variation in the outcome was attributable to the relationship with the covariates. In this section, we present selected additional simulation results for scenarios 1, 2, and 3 for which approximately 20% or approximately 5% of the variation in the outcome was attributable to the relationship with the covariates.

| Feature Selection | Model Fit | % Selection Relevant (1+) | Avg # Relevant Var | % Other Var (1+) | Avg # Additional Var | Avg % Variation Unexplained by Pred. |
|---|---|---|---|---|---|---|
| GA: MSE | LR | 85.0 | 1.72 | 100 | 16.0 | 110.6 |
| GA: BIC | LR | 13.0 | 0.13 | 87 | 0.87 | 105.1 |
| GA: MSE | SVR | 100 | 4.99 | 70 | 1.94 | 21.4 |
| GA: BIC | SVR | 100 | 4.96 | 3.0 | 0.03 | 14.8 |
| RFE | SVR | 100 | 4.42 | 100 | 3.59 | 52.9 |
| LASSO | LR | 0 | 0 | 0 | 0 | 104.8 |
| None | RF | 100 | 5 | 100 | 45 | 31.3 |
| None | SVR | 100 | 5 | 100 | 45 | 95.5 |

Table A9: Results of simulation 1 setting with 50 total covariates X, 5 of them truly associated with the outcome Y with an entirely quadratic relationship, with 20% of the variance in Y attributable to relationship with X.

| Feature Selection | Model Fit | % Selection Relevant (1+) | Avg # Relevant Var | % Other Var (1+) | Avg # Additional Var | Avg % Variation Unexplained by Pred. |
|---|---|---|---|---|---|---|
| GA: MSE | LR | 100 | 5 | 100 | 15.81 | 8.74 |
| GA: BIC | LR | 100 | 5 | 0 | 0 | 4.9 |
| GA: MSE | SVR | 100 | 5 | 100 | 13.65 | 11.1 |
| GA: BIC | SVR | 100 | 5 | 0.01 | 0.01 | 6.66 |
| RFE | SVR | 100 | 5 | 100 | 3 | 37.34 |
| LASSO | LR | 100 | 5 | 0.03 | 0.03 | 4.88 |
| None | RF | 100 | 5 | 100 | 45 | 22.56 |
| None | SVR | 100 | 5 | 100 | 45 | 20.01 |

Table A10: Results of simulation 1 setting with 50 total covariates X, 5 of them truly associated with the outcome Y with an entirely linear relationship, with 20% of the variance in Y attributable to relationship with X.

| Feature Selection | Model Fit | % Selection Relevant (1+) | Avg # Relevant Var | % Other Var (1+) | Avg # Additional Var | Avg % Variation Unexplained by Pred. |
|---|---|---|---|---|---|---|
| GA: MSE | LR | 100 | 4.97 | 100 | 15.9 | 61.8 |
| GA: BIC | LR | 100 | 4.79 | 0 | 0 | 58.7 |
| GA: MSE | SVR | 100 | 4.99 | 99.2 | 8.3 | 44.7 |
| GA: BIC | SVR | 100 | 4.98 | 3.6 | 0.036 | 38.2 |
| RFE | SVR | 100 | 4.98 | 100 | 3.02 | 44.7 |
| LASSO | LR | 100 | 5.0 | 3.4 | 0.04 | 56.7 |
| None | RF | 100 | 5 | 100 | 45 | 37.0 |
| None | SVR | 100 | 5 | 100 | 45 | 69.3 |

Table A11: Results of simulation 2 setting with 50 total covariates X, 5 of them truly associated with the outcome Y including an underlying interaction relationship, with 20% of the variance in Y attributable to relationship with X.

## A.4 Kidney transplant registry descriptive characteristics

Methods and results of the data application are presented in Section 2.5 of the full text. The 32 covariates used in analysis for the 21,121 patients used in analysis are

| Feature Selection | Model Fit | % Selection Relevant (1+) | Avg # Relevant Var | % Other Var (1+) | Avg # Additional Var | Avg % Variation Unexplained by Pred. |
|---|---|---|---|---|---|---|
| GA: MSE | LR | 100 | 14.83 | 100 | 72.93 | 37.63 |
| GA: BIC | LR | 100 | 14.75 | 57.7 | 0.84 | 29.62 |
| GA: MSE | SVR | 100 | 14.98 | 100 | 33.61 | 28.22 |
| GA: BIC | SVR | 100 | 14.98 | 85.0 | 1.70 | 20.95 |
| LASSO | LR | 100 | 14.96 | 95.4 | 6.03 | 29.49 |
| None | RF | 100 | 15 | 100 | 285 | 37.87 |
| None | SVR | 100 | 15 | 100 | 285 | 50.60 |

Table A12: Results of simulation 3 setting with 50 total covariates X, 5 of them truly associated with the outcome Y and correlated with non-associated covariates, with 20% of the variance in Y attributable to relationship with X.

detailed in Table A13 below. MICE was used to impute missing data for the following variables (percent missing displayed): health insurance (1.2%), education (2.6%), functional status (1.4%), and diastolic blood pressure (0.5%).

| **Categorical Donor Characteristics** | N (%) |
|---|---|
| Gender | |
| Female | 13,629 (64.5) |
| Male | 7,492 (35.5) |
| Race | |
| Black | 1,829 (8.7) |
| Other | 1,196 (5.7) |
| White | 18,096 (85.7) |
| Ethnicity | |
| Latino | 3,064 (14.5) |
| Non-latino or unknown | 18,057 (85.5) |
| Blood type | |
| A | 4,869 (23.1) |
| AB | 208 (1.0) |
| B | 1,654 (7.8) |
| O | 13,283 (62.9) |
| Other | 1,107 (5.2) |
| Citizenship | |
| Yes | 20,039 (94.9) |

| | |
|---|---|
| No | 1,082 (5.1) |
| Education | |
| Did not attend college | 4,881 (23.1) |
| Attended college | 16,240 (76.9) |
| Has health insurance | |
| Yes | 19,677 (93.2) |
| No | 1,444 (6.8) |
| History of cancer | |
| Yes | 564 (2.7) |
| No | 20,557 (97.3) |
| History of hypertension | |
| Yes | 989 (4.7) |
| No | 20,132 (95.3) |
| History of cigarette use | |
| Yes | 5,173 (24.5) |
| No | 15,948 (75.5) |
| Other tobacco use | |
| Yes | 716 (3.4) |
| No | 20,230 (95.8) |
| Unknown | 175 (0.8) |
| Relationship to organ recipient | |
| Biological | 8,939 (42.3) |
| Partner | 2,866 (13.6) |
| Unrelated | 9,316 (44.1) |

| **Continuous Donor Characteristics** | Mean (sd) |
|---|---|
| Age (years) | 44.26 (12.33) |
| Height (cm) | 168.56 (9.96) |
| Weight (kg) | 76.66 (14.9) |
| BMI | 26.88 (4.11) |
| Pre-transplant eGFR | 97.13 (17.22) |
| Preoperative diastolic blood pressure | 74.27 (9.29) |

Table A13: The mean and standard deviation (if continuous) or frequency and proportion (if categorical) of each covariate are presented for the 21,121 kidney donors used in the analysis.

## A.5 Alternate kernel simulation

Simulations from Section 2.4 of the full text were replicated in a sensitivity analysis to examine the impact of an alternate choice of kernel for the support vector regression (SVR) model. While the radial-basis kernel was used in the full text, we apply a linear kernel here. Outside of kernel choice, all other simulation methods were consistent with the simulations reported in Section 2.4 of the full text.

Results under a quadratic data generation scheme, scenario 1 in the text, with $p = 50$ covariates, $\pi = 0.1$ proportion of them related to the outcome, are presented and can be compared to results presented in the text in Figure 2 and in the Appendix Tables A1-A4.

Changing the kernel in our method yielded results similar to using a linear regression in combination with a genetic algorithm feature selection, in all scenarios assessed. Performance was better under scenarios with mostly linear relationships between the outcomes and covariates, as would be expected under a linear kernel.

| Ref. Table | Feature Selection | Model Fit | % Selection Relevant (1+) | Avg # Relevant Var | % Other Var (1+) | Avg # Additional Var | Avg % Variation Unexplained by Pred. |
|---|---|---|---|---|---|---|---|
| A1 | GA: MSE | SVR | 85.2 | 1.60 | 100 | 13.89 | 108 |
| A1 | GA: BIC | SVR | 14.8 | 0.15 | 85.4 | 0.86 | 105 |
| A2 | GA: MSE | SVR | 100 | 4.97 | 100 | 13.45 | 73.5 |
| A2 | GA: BIC | SVR | 100 | 4.93 | 19.2 | 0.22 | 71.3 |
| A3 | GA: MSE | SVR | 100 | 5.0 | 100 | 13.42 | 37.5 |
| A3 | GA: BIC | SVR | 100 | 5.0 | 20.6 | 0.23 | 35.4 |
| A4 | GA: MSE | SVR | 100 | 5.0 | 100 | 13.48 | 3.5 |
| A4 | GA: BIC | SVR | 100 | 5.0 | 19.0 | 0.21 | 1.9 |

Table A14: Results of simulation scenario 1 with 50 total covariates, 5 of them truly associated with the outcome Y. Underlying quadratic relationship decreases as underlying linear relationship increases moving down the table. Results are all using the linear kernel for SVR model and can be compared to the noted tables for performance with the RBF kernel.

| Ref. Table | Feature Selection | Model Fit | % Selection Relevant (1+) | Avg # Relevant Var | % Other Var (1+) | Avg # Additional Var | Avg % Variation Unexplained by Pred. |
|---|---|---|---|---|---|---|---|
| 1 | GA: MSE | SVR | 100 | 5 | 100 | 13.48 | 55.97 |
| 1 | GA: BIC | SVR | 100 | 5 | 16.4 | 17.4 | 53.72 |

Table A15: Results of simulation scenario 2 with 50 total covariates, 5 of them truly associated with the outcome Y, including underlying interaction relationships not supplied to the model. Results are using the linear kernel for SVR model and can be compared to the noted tables for performance with the RBF kernel.

| Ref. Table | Feature Selection | Model Fit | % Selection Relevant (1+) | Avg # Relevant Var | % Other Var (1+) | Avg # Additional Var | Avg % Variation Unexplained by Pred. |
|---|---|---|---|---|---|---|---|
| 2 | GA: MSE | SVR | 100 | 5 | 100 | 12.86 | 3.36 |
| 2 | GA: BIC | SVR | 100 | 5 | 18.6 | 0.19 | 1.82 |
| 3 | GA: MSE | SVR | 100 | 10 | 100 | 11.29 | 3.70 |
| 3 | GA: BIC | SVR | 100 | 9.99 | 0.8 | 0.01 | 2.43 |

Table A16: Results of simulation 3 setting with 50 total covariates $\mathbf{X}$. Rows 1 and 2 refer to 5 covariates truly associated with the outcome $\mathbf{Y}$ and each correlated to 9 variables unrelated to the outcome with AR-1 correlation structure; rows 3 and 4 refer to 10 covariates truly related to the outcome and correlated with each other with AR-1 block correlation structure. Results are using the linear kernel for SVR model and can be compared to the noted tables for performance with the RBF kernel.