Noname manuscript No.
(will be inserted by the editor)

Outcomes Matter: Estimating Pre-Transplant Survival Rates of Kidney-Transplant Patients Using Simulator-Based Propensity Scores

Inbal Yahav · Galit Shmueli

Received: June 30, 2011 / Accepted: date

Abstract The current kidney allocation system in the United States fails to match donors and recipients well. In an effort to improve the allocation system, the United Network of Organ Sharing (UNOS) defined factors that should determine a new allocation policy, and particularly patients' potential remaining lifetime without a transplant (pre-transplant survival rates). Estimating pre-transplant survival rates is challenging because the data available on candidates and organ recipients is already "contaminated" by the current allocation policy. In particular, the selection of patients who are offered (and decide to accept) a kidney is not random. We therefore expect differences in mortality-related characteristics of organ recipients and of candidates who have not received transplant. Such differences introduce bias into survival models.

Existing approaches for tackling this selection bias either ignore the difference between candidates and recipients or assume that selection to transplant is based solely on patients' pre-transplant information, irrespective of the potential allocation outcome. We argue that in practice the allocation is dependent on the anticipated allocation outcome, which is a major factor in selection to transplant. Moreover, we show that ignoring the anticipated outcome increases model bias rather than decreases it. In this paper, we propose a

This work was supported in part by Health Resources and Services Administration contract HHS/HRSA SRTR. The content is the responsibility of the authors alone and does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

Inbal Yahav
Bar Ilan University
Graduate School of Business
Tel.: +972-3-5318913
E-mail: inbal.yahav@biu.ac.il
Galit Shmueli
Indian School of Business
E-mail: galit shmueli@isb.edu

novel simulator-based approach (SimBa) that adjusts for the selection bias by taking into account both pre-transplant and anticipated outcome information using simulation. We then fit survival models to kidney transplant waitlist data and compare the different adjustment methods. We find that SimBa not only fits the data best, but also captures a key aspect of the current allocation policy, namely, that the probability of kidney allocation increases in the expected pre-transplant life years.

Keywords Selection bias · Pre-transplant survival rate · Kidney allocation · Propensity scores · Survival analysis · Simulation

1. Introduction

According to the Scientific Registry of Transplant Recipients (SRTR) annual statistics, more than 90,000 candidates with kidney failure End-Stage Renal Disease (ESRD) are currently waiting for transplantation in the United States. Whereas the number of annual transplants stands at approximately 13,500, the number of annual waiting list additions reaches 30,000. This imbalance between demand and supply of organs raises a need for an efficient organ allocation policy to determine the order in which candidates are offered an organ, when one becomes available.

Under the current kidney allocation system in the United States, kidneys are allocated to patients primarily through a combination of tissue matching, sensitization level (the level of sensitization to donor antigens, measured by Panel Reactive Antibody), and waiting time. However, due to recent trends in medicine and the shortfall of kidney supply, the current system fails to match donors and recipients well. Ideally, kidney allocation should take into account the potential outcome in terms of transplant success, post-transplant lifetime, etc. In an effort to improve the allocation system, the United Network of Organ Sharing (UNOS) defined factors that should determine a new allocation policy, a major factor being patients' potential remaining lifetime without a transplant.

Survival rates of patients with kidney failure ESRD provide important medical information that affects candidates' transplant options, and therefore the development of an efficient kidney allocation policy. *Pre-Transplant Life Years* (*PTLY*) refers to the lifetime of an ESRD patient who has not received a transplant. PTLY models are useful for various purposes including comparing candidates' survival rates under different policies, estimating overall waitlist mortality rates, and measuring lifetime gained from a specific transplant (called *Life Years From Transplant (LYFT)*). The new allocation policy called KAS (Kidney Allocation Scores; OPTN/UNOS (2008)) that will be replacing the existing policy, will rely on LYFT for allocation decisions.

Consider the two types of waitlist patients: *recipients* who receive a transplant during the study period, and *candidates* who have not received a kidney. Modeling patients' PTLY using transplant waitlist data poses several statistical challenges. First, the PTLY for a kidney recipient is right-censored at the

time of transplant, and therefore we have no complete information on recipients' PTLY values. That is, their life years had they not received transplant. PTLY is also right-censored for candidates who remained alive during the study period, where the end of the study period creates right censoring. To build a survival model we must therefore consider which data to use. While PTLY values of deceased and live candidates can be modeled using survival models (due to the non-informative censoring), integrating recipients' data into such models is a challenge due to the informative censoring resulting from the non-random nature of kidney allocation decisions. The major challenge is that existing waitlist data are 'contaminated' by the existing allocation policy, under which patients are not selected for transplant at random. Therefore, transplant recipients can greatly differ from candidates in terms of their PTLY-related characteristics. These differences can potentially introduce bias into survival rate estimates. Moreover, when patients are offered a kidney, they very often refuse it in the hope of obtaining a more suitable kidney while they are high on the priority list. This extra step further differentiates recipients from candidates.

There are several approaches to tackling the effects of the allocation policy on waitlist data, and the resulting informative censoring. One approach, called *complete-case analysis* (CCA), assumes statistical equivalence of candidates and recipients in terms of PTLY-related characteristics (see e.g., SRTR (2007a,b); Wolfe et al (2008)). Recipients, who have right-censored PTLY values, are then excluded from the sample and candidates are assumed to be representative of the population of all ESRD patients. CCA has two disadvantages:

1. The loss of efficiency due to excluding data on recipients and using a smaller sample, and
2. Inconsistent estimates of survival rates when the underlying censoring mechanism (the allocation policy) depends on the outcome.

A second approach is to impute missing recipients' PTLY values by matching recipients with candidates who have similar medical histories. One challenge is defining "similar". Another challenge is that the medical history of patients is only recorded when they join or leave the list, but is not updated throughout the long period between joining and leaving except for few sporadic updates for different patients at different times. Methods based on data mining algorithms and statistical matching have been proposed.

A third approach aims at adjusting the selection bias by reweighting the PTLY value for each candidate by his/her transplant probability, thereby creating a representative sample of ESRD patients. Weights are traditionality estimated using logistic regression or discriminant analysis, yielding "propensity scores" (Rosenbaum and Rubin, 1983). The challenge with using propensity scores is that they rely on the assumption of *strong ignorability*. In the context of kidney transplants, strong ignorability implies that the assignment to transplant, conditional on observed pre-transplant covariates, is indepen-

dent of the potential transplant outcome. Such independence in this context is highly unlikely.

In this paper, we propose a method for reducing selection bias for estimating PTLY survival rates using a novel approach. Our method generates improved weights that do not suffer from the weaknesses of existing methods. We argue that the medical outcome of a transplant along with other sources of uncertainty, such as organ acceptance decisions, are major factors that introduce bias into PTLY survival models. Hence, the assumptions underlying CCA, propensity scores, and other existing methods are violated. We propose a simulator-based model, called *SimBa*, which computes weights based not only on patients' selection to transplant covariates but also on their simulated outcome. Simulated outcomes are generated using the computer simulation program developed by SRTR (operated by the non-profit organization UNOS). The simulation approach allows us to meet the assumption of strong ignorability and to compute propensity scores based on patients' observed covariates as well as on unobserved covariates that correlate with the outcome.

We compare SimBa to CCA and standard propensity score approaches by applying them to a large waitlist dataset for kidney transplants in the United States between 2000-2010 (see Appendix A for a description of the variables in this dataset). We find that SimBa not only outperforms the other methods in terms of model fit, but more importantly it captures a crucial and surprising aspect of the current allocation policy: *the probability of transplant increases in expected PTLY*. While the current policy prioritizes high-risk patients, the resulting allocation actually gives an advantage to patients with anticipated high pre-transplant life years. In other words: outcomes matter!

The contribution of our paper is both methodological and practical. Methodologically, we introduce an improved approach for estimating pre-transplant survival rates in the presence of an existing allocation policy. Applying our approach to real data and comparing it to existing methods, we find that other bias-correcting methods actually introduce more bias. Our proposed method contributes to practice by highlighting that with the current allocation policy, the probability of receiving a kidney increases in anticipated life years.

The remainder of the paper is organized as following. We introduce terminology and notation in Section 2. Section 3 describes existing approaches for addressing recipients' missing PTLY values. Section 4 introduces our simulator-based approach and its advantages. In Section 5 we describe a survival model and how SimBa propensity scores are incorporated into the model. In Section 6 we apply the different methods to the 2000-2010 waitlist dataset for kidney transplant and discuss and compare the different resulting survival models. We discuss the main results and draw conclusions in Section 7.

2. Terminology and Notation

Waitlist data contain patients of three types:

Recipients are patients who received a transplant during the study period.

Live candidates are patients who did not receive a transplant during the study period and remained alive by the study end.

Deceased candidates are patients who did not receive a transplant during the study period and died during the study period.

We use $PTLY$ for Pre-Transplant Life Years. This is measured by the number of years from joining the waitlist until death, given that the patient did not receive transplant.

In the remainder of the paper we use the following notation:

$PTLY_i$ = pre-transplant life years for waitlist patient i.

p_i = Probability of transplant for waitlist patient i.

$PS_i = \hat{p}_i$ = Propensity score for patient i, an estimate of p_i.

3. Existing Approaches for Handling Recipients' Missing PTLY Values

We present a brief survey of the kidney transplant and related literature regarding approaches to handling missing PTLY values for kidney recipients, for the purpose of modeling the survival rates of ESRD patients. The three approaches are: excluding data on recipients, imputing recipients' data, and reweighting candidates' data. Methods include complete case analysis; data imputation via logistic regression and classification and regression trees; and reweighting via propensity scores using logistic regression.

3.1 Excluding Missing Data: Complete Case Analysis (CCA)

Complete Case Analysis (CCA) is an approach for handling data with missing values by using only "complete" observations and dropping all observations with missing values. In the context of modeling the PTLY of ESRD patients using data from the UNOS waitlist, all kidney recipients have right-censored PTLY values, where censoring is informative. That is, we do not know how long they would have survived had they not received transplant. Hence, the CCA approach treats recipients' PTLY values as missing values and drops all kidney recipients from the modeling.

CCA is based on the assumption that excluded and included observations are statistically equivalent in terms of mortality-rate-related characteristics (SRTR, 2007a,b; Wolfe et al, 2008; Schaubel et al, 2009). The main limitation of the CCA approach in our context is that in practice the underlying assumption is violated: recipients and candidates differ statistically in terms of their medical conditions which are related to their mortality rate. Figure 1 shows such differences, comparing a sample of candidates and recipients from the UNOS waitlist. It shows the differences in terms of diabetes status ("DIAB"),

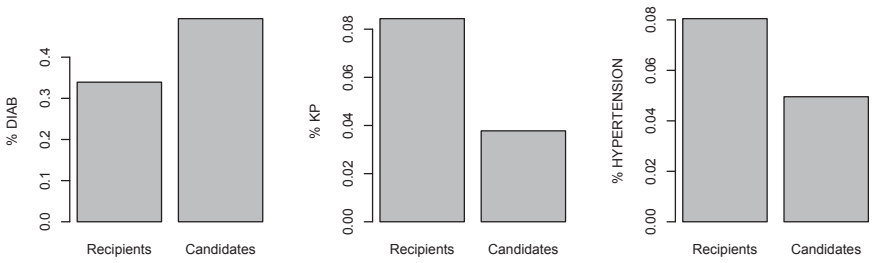

Fig. 1 Comparison of kidney recipients and candidates (who have not received transplant), in terms of medical conditions.

need for a simultaneous kidney-pancreas transplant ("KP"), and hypertension information ("HYPERTENSION").

Not only does the allocation policy affects the outcome; in practice, many patients (45%, according to Zenios (2004)) choose to refuse an offered organ that is not best suited to them, as advised by their physicians. Hence, although the allocation system does not prioritize patients according to their PTLY (with the exception of urgent cases), patients' choice to accept or reject an organ is also outcome-dependent.

The differences between the two patient groups mean that using only candidates' data for modeling survival rates of all ESRD patients will result not only in loss of efficiency due to the reduced sample size, but more importantly it will introduce bias in the estimated model and therefore might lead to invalid inference. The faults of CCA and its substantial bias when missingness is dependent on the outcome are well-known (e.g., Demissie et al (2003)).

3.2 Data Imputation

Data imputation is a common alternative to CCA that is superior in terms of bias reduction, both theoretically and in various applications. The idea is to impute the missing PTLY data for recipients.

Wolfe (2007) proposed a conceptual procedure for imputing a recipient's PTLY by matching his/her medical conditions *at the time of kidney offer* to a "similar" candidate in the waitlist. However, the concept of "medical similarity" remains undefined, and the method was not implemented in practice on real data.

Yahav and Shmueli (2010) introduced a two-step method for imputing recipients' PTLY value. In the first step, death incidences of recipients are imputed using a classification tree procedure that utilizes candidates' medical history as inputs and their death as a binary output variable. Then, *conditional on death incidence*, a recipient's PTLY is imputed using a regression tree.

The challenge with both imputation approaches is that they require patients' medical history during the entire period from arrival to the waitlist onwards. The medical history obviously changes over time, and therefore must be updated. Yet, the UNOS dataset includes only the medical history recorded at arrival time, departure time (such as transplant, if applicable), and at main status change periods (e.g., the patient moves to another center or the patient's pancreas failed while waiting for a kidney). The actual time-changing medical history is not recorded. Therefore, recipients can be matched with candidates mainly based on their initial record, and not based on their health progression over time.

A third imputation approach, propensity score matching (Rosenbaum and Rubin, 1983), relies on using propensity scores for creating a matched sample of recipients and candidates. Propensity scores is a data-driven approach to studying treatment effects when the investigator has no control over the treatment assignment. Each observation is assigned a score based on its conditional probability of being selected to treatment (in this case, transplant), given the individual's observed covariates. Propensity score models are commonly estimated by fitting a logistic regression to the entire study population (Stürmer et al, 2005), and then using the model to predict selection scores. In the context of data imputation, propensity scores can be used to create a matched sample of candidates and recipients, so that the two groups are similar in terms of their medical history covariates. The matched samples are then used for imputing the missing PTLY values for recipients.

The main weakness of using propensity scores and estimating transplant probabilities using logistic regression (or similar statistical models such as discriminant analysis) is the underlying assumption of *strong ignorability*. This assumption implies that the assignment to transplant, conditional on observed pre-transplant covariates, is independent of the potential transplant outcome. Such independence is highly unlikely because in practice kidney allocation is not random and depends on the expected outcome of the transplant. Patients are offered an organ, and decide to accept it, if both the policy planner and the patient (and his/her physician) believe that the transplant will be successful in terms of significantly increasing the patient's lifetime and quality of life. Hence, the selection to transplant is not only a function of the patient's observed medical covariates, but also of the potential outcome and the patient's acceptance decision, which are not captured by the logistic model.

A second weakness of propensity scores imputation is that it does not take into account the censoring effect of the study period. Ignoring the right-censoring of PTLY values due to waitlist departure (mostly due to death) or the truncated study period, introduces bias. In particular, patients with low PTLY values due to censoring should have a smaller transplant probability within the study period, compared to patients with the same health conditions (upon arrival to the waitlist) who arrived earlier. Yet, the propensity scores approach would assign these two types of patients an equal transplant probability.

Lastly, like imputation, standard propensity score imputation procedures do not account for sporadic medical health updates such as those available in UNOS waitlist. Recall that information is only recorded upon arrival to the waitlist and at few special events such as death or change of insurance plan. Hence, medical updates are only available in unstructured form (different updates for different patients at different times). Yet such updates can be informative.

3.3 Data Reweighting

Reweighting data according to selection probability is common in survey analysis for reducing selection bias, when some sub-populations have a lower probability of being selected to the sample. Weights are used to adjust for deviations between the variable distribution in the sample and the target population (Henry, 1990). Unlike post-stratification weighting, where only a small number of categorical covariates are used for computing the weights, propensity scores allow the inclusion of multiple numerical and categorical covariates.

Another difference between the contexts of surveys and kidney allocation is knowledge about the covariates' distribution in the population and the relation between the sample and the population. Such knowledge allows direct estimation of weights. For example, Binder (1992) and Lin (2000) use weighting to reduce bias due to survey non-response, assuming that the covariate distribution in the population is known (e.g., census data) and the sample design is controlled by the researcher. In contrast, in the kidney allocation context the covariate distribution in the population of ESRD patients is unknown and the allocation of kidneys is not under the control of the researcher.

Propensity scores are widely used to account for selection bias in medical research (D'Agostino, 1998). For example, Cho et al (2006) study the mortality rate of patients treated with different dialysis treatments after acute kidney injury. The authors use propensity scores to account for selection to different treatment options; Polkinghorne et al (2004), D'Agostino (2007) and Bavaria et al (2007) apply propensity scores to vascular access and alternative treatment options and their effect on patients' survival. Engoren et al (2002) examine the effect of blood transfusion after cardiac operation on long-term survival. Selection to transfusion is adjusted with propensity scores. Common to all these studies is that both treatment and non-treatment groups include censored and uncensored survival rates. In contrast, in the context of kidney allocation, the recipient group contains no uncensored survival rates. This special data structure means that propensity score weights are applied only to candidates' PTLY values, and the weighted sample of candidates can then be considered to represent all ESRD patients.

Propensity scores are useful even when the population covariates' distribution is unknown. Pan and Schaubel (2008, 2009) used logistic regression to compute propensity scores in the context of post-transplant *graft lifetime* (the lifetime of a kidney after it is transplanted). There, an observation is a graft

and the probability of selection is the probability of being transplanted, as opposed to being discarded.

To the best of our knowledge, this approach has not been used to estimate pre-transplant survival rates. As with the use of propensity scores for creating a matched sample of candidate-recipient pairs (see Section 3.2), the three limitations of assuming strong ignorability, ignoring censoring, and excluding medical updates during the study period lead to biased results.

4. SimBa: Simulator-Based Propensity Scores

We propose a simulator-based solution, called *SimBa*, for computing propensity scores of waitlist patients. SimBa overcomes the limitations of existing approaches, namely violating the assumption of strong ignorability, ignoring censoring, and excluding medical updates during the study period that are in unstructured form.

Recognizing that allocation outcomes are detrimental to allocation itself, SimBa accounts for information about potential allocation outcomes that are unavailable in waitlist data in a probabilistic yet realistic fashion. For example, SimBa accounts for patients' decisions to accept an offered kidney, and for possible transplant outcomes such as death, success or failure of transplant, and re-listing. SimBa also accounts for waitlist dynamics such as changes in patients' health condition. The SimBa approach therefore meets the assumption of strong ignorability that underlies propensity scores, as well as accounts for the changing nature of patient's health condition.

SimBa uses simulation in two forms: First, it uses the Kidney-Pancreas Simulated Allocation Model (KPSAM) simulator, developed by SRTR, for simulating the allocation of organs to waitlist candidates according to the current Priority Points allocation policy[1]. To the best of our knowledge, there has been no attempt in the literature to integrate real waitlist data with simulation allocation data using the realistic SRTR simulator for estimating pre-transplant survival rates. Second, to incorporate the uncertainty associated with the KPSAM simulator so that results can be generalized beyond a particular waitlist dataset, SimBa simulates multiple allocations by randomizing donor arrivals. Next, we discuss the two simulation components in further detail.

[1] Health condition in KPSAM is modeled through (1) medical updates recorded during major status changes (given to the simulator as an input), and (2) time-dependent variables, such as age, dialysis time, etc. that the simulator automatically keeps updated over time. The simulation also incorporates health updates in a more realistic way compared to any static model (such as the logistic regression): updated information is introduced into the model only when it becomes available (i.e., when the change occurs), and only if it occurs before the event of transplant or death. In contrast, static models include status update information in a static fashion, as if it is available upon patient arrival, and irrespective of the outcome.

4.1 Kidney-Pancreas Simulated Allocation Model (KPSAM) Simulator

The Scientific Registry of Transplant Recipients (SRTR), under direction of the U.S. Health Resources and Services Administration (HRSA), developed the Kidney-Pancreas Simulated Allocation Model (KPSAM) simulator version 4.2 (KPSAM, 2009) for simulating the allocation of organs to waitlist candidates according to the current Priority Points allocation policy (for more information on the current policy see UNOS (2011)). The simulator takes into account patients' decisions to accept an offered kidney, transplant outcomes such as death, success or failure of transplant, patient re-listings, as well as changes in patients' health condition. It then simulates the allocation of organs to waitlist candidates.

The KPSAM simulator is based on an event-sequenced Monte Carlo technique, where some of the modeled processes (organ acceptance, relisting, post-graft survival, and non-relist death) are stochastic. The simulator samples pseudo-random numbers to generate a realization of these processes over the specified time period. Choosing different initial random seeds therefore allows generating different realizations from these stochastic models.

Recognizing the insufficiency of freely available waitlist data, and the value of KPSAM for estimating pre-treatment survival rates, we purchased rights to use the simulator program for a period of one year.

We apply KPSAM to an existing waitlist, by applying the current policy and using multiple initial random seeds to capture the variability associated with the stochastic processes that KPSAM models. We then take an extra simulation step (described in the next section) to allow generalization beyond the time-sequenced organ and patient arrivals in an existing waitlist. Our purpose is to estimate the probability of transplant for each patient in the waitlist (p_i) under the current policy. The estimated probabilities are then used as propensity scores.

4.2 Simulating Allocation Replications and Computing Propensity Scores

Computing propensity scores requires estimating each patient's conditional probability of receiving transplant, that is, being offered a kidney and accepting it under the current allocation policy. To that end, we generate replications of the waitlist where *donor arrivals are randomized*, and then the allocation is simulated using the KPSAM simulator, which takes into account the outcomes. The allocation policy heavily relies on the order of donor arrivals, due to its local maximization fashion (a kidney is allocated to the *current* most suitable patient). Therefore, randomizing the order of donor arrivals (or equivalently, kidney arrivals) allows us to generalize results beyond the particular organ arrival order in the study period. Because the demand for kidneys is much larger than supply, various scenarios can occur within a study period that would result in estimated PS_i values of 0 or 1 which do not reflect the actual allocation in the longer run. For example, patients who wait for a rare

kidney type and join the waitlist near the end of the study period might have $PS_i = 0$ if one considers only the actual donor arrivals, but a slightly higher probability under a different arrival order. Another scenario is a high-priority patient with low sensitization level who competes with a low-priority patient with high sensitization level. According to the allocation policy, the next available organ will be assigned to the former, regardless of its compatibility with the latter patient, implying that the low-priority patient's p_i depends on the order of compatible kidney arrivals. Randomizing donor arrivals can overcome such limitations. Other scenarios that depend on the organ arrival include (1) changes in patients' health condition and consequently, their type of required kidney and/or their waitlist priority, and (2) pre-transplant departures from the waitlist for reasons of death, living-donor donations, transfer to another center, etc.

The SimBa algorithm repeatedly simulates the allocation of organs to waitlist candidates using KPSAM, by generating M bootstrap samples so that donor arrivals are randomized. For each of the M bootstrap samples, we simulate N samples using different initial random seeds in the KPSAM simulator, resulting in a total of $M \times N$ samples. Given a set of replications, SimBa computes the "success probability" for each candidate. This probability of receiving transplant (that is, being offered and accepting an organ) is estimated using the proportion of times out of $M \times N$ that the patient "received" transplant.

SimBa uses three types of datasets for generating propensity scores from the resampled data:

1. The list of waitlist patients and their covariates upon arrival, as given at the start of the study period. In our case, the start date is 1/1/2000 and the covariates are described in Appendix A.
2. Arrivals and departures of patients (due to transplant, death, or other reasons) and any major status changes during the study period. In our case the study period is 2000-2010.
3. Simulated donor arrival times during the study period. We generate a total of M donor datasets.

The data used for generating these three types of datasets in our application are described in detail in Section 6.1.

To simulate donor arrival times, we first decompose the actual donor information into *arrival timestamp* and *donor information* (age, blood type, cause of death, etc.). We then generate a new donor dataset of the same size as the original dataset by using a bootstrap sample of donors' information (Efron and Tibshirani (1993)) to randomize the order in which they arrive. Finally, we recombine the *actual* arrival timestamp with the *sampled* donor information. This means that the same donors arrive during our study period, but in a different order. The simulated donors dataset maintains the characteristics of the original arrival information in terms of both organ arrival rate and organ distribution. Figure 2 illustrates this process.

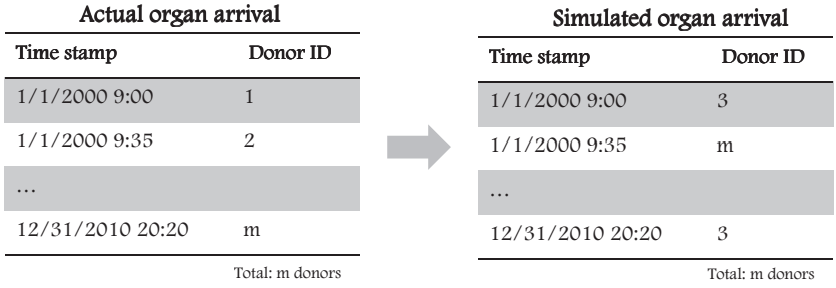

Fig. 2 Illustration of Donors' Simulated Arrival Data

The propensity score for patient i (PS_i) is the proportion of times the patient received transplant according to our algorithm. For example, if a patient received an organ k times out of the $M \times N$ runs, his/her estimated transplant probability would be $PS_i = k/(M \times N)$. A schematic representation of the SimBa algorithm is given in Figure 3.

The novelty of SimBa is its ability to meet the assumption of strong ignorability by basing the propensity scores on both observed and unobserved patients' covariates, such as their acceptance decision, and on the actual potential outcome (death, success/failure of transplant, candidates' relisting, etc.). Furthermore, by randomizing donor arrivals, SimBa allows computing the mean transplant probability given a probability distribution of organ types, thereby avoiding overfitting the probabilities to a particular waitlist.

4.3 From Propensity Scores to Weights

Once propensity scores are computed via SimBa, logistic regression, or any other method, we use the binning approach by Rosenbaum and Rubin (1984) to create selection-bias correction weights. Weighting each of the candidates' PTLY values is intended to generate a representative sample of the complete waitlist (candidates and recipients). The procedure is:

1. Sort patients by their propensity scores
2. Create J homogeneous subgroups of patients with similar propensity scores
3. The weight for a candidate in subgroup j is computed by
   $w_j = \frac{\#\text{patients in subgroup j}}{\#\text{candidates in subgroup j}}$

5. Estimating Pre-Transplant Survival Rates Using a Weighted Accelerated Failure Time (AFT) Model

We use the parametric Accelerated Failure Time (AFT) model to estimate ESRD patients' survival curves. Compared to the popular semi-parametric Cox

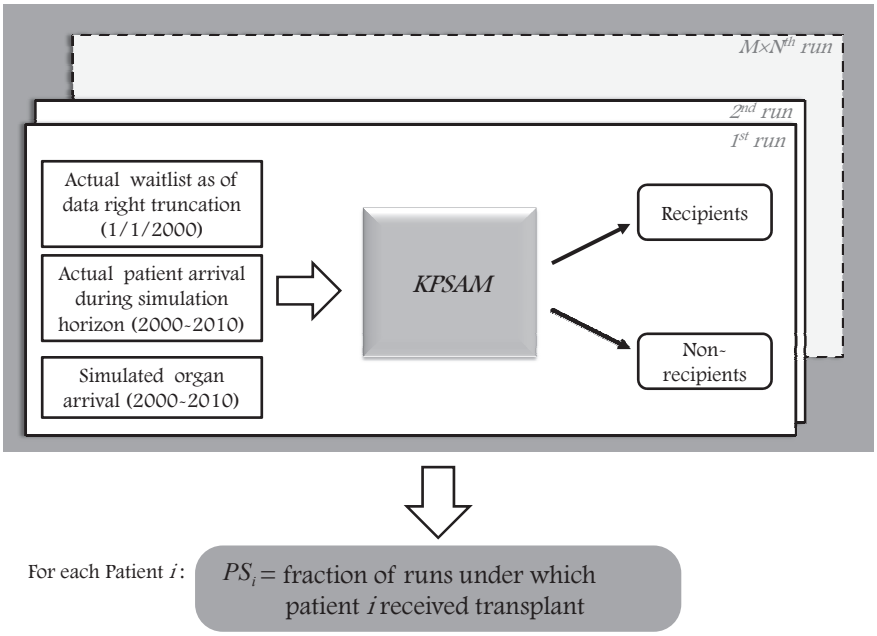

Fig. 3 Schematic representation of SimBa.

proportional hazards model (see, e.g., Wolfe et al (2008)), in our context the parametric AFT model offers two important advantages. First, AFT models are known to be more robust to unmeasured or neglected covariates (also referred to as *hidden heterogeneity*), compared to the Cox model (Shyur et al (1999); Lambert et al (2004)). Second, unlike the Cox model that does not allow extrapolation beyond the study period, the AFT model enables us to extrapolate survival rates beyond the length of the study period. Because the study period is typically shorter than many waiting times, it is important to be able to estimate longer survival rates.

Following Yahav and Shmueli (2010), we use an AFT model with a Weibull distribution, which offers sufficient flexibility and better model fit compared to other distributions. The Weibull AFT model with covariates $\mathbf{z}$ models the survival at time t as:

$$S(t|\mathbf{z}) = S_0(t)e^{e^{\boldsymbol{\gamma}'\mathbf{z}}} = e^{-t^{\alpha}e^{\boldsymbol{\gamma}'\mathbf{z}}}, \quad t > 0, \alpha > 0 \tag{1}$$

where $S_0(t) = \exp(-t^{\alpha})$ is the baseline Weibull survival probability, α is the shape parameter, $\boldsymbol{\gamma}$ is a vector of parameters and $\mathbf{z}$ is the vector of covariates. This model essentially places individuals with different covariates on different time scales. The Weibull AFT model can also be parameterized as a proportional hazards model (see, e.g., Kleinbaum and Klein (1995)), so that the hazard is given by

$$h(t|\boldsymbol{z}) = \alpha t^{\alpha-1}e^{\boldsymbol{\gamma}'\mathbf{z}}, \tag{2}$$

thereby assuming a linear relationship between the log of failure time ($\log(T)$) and the covariates ($\mathbf{z}$), with an error term (ε) that follows a Weibull distribution:

$$\log(T) = \alpha + \boldsymbol{\gamma}^{'}\mathbf{z} + \varepsilon. \quad (3)$$

Estimating model parameters α and $\boldsymbol{\gamma}$ can be achieved using ordinary maximum likelihood estimation, or, in the presence of selection bias, via weighted maximum likelihood estimation (Field and Smith (1994)). We take the latter approach, using weights derived from the SimBa propensity scores described in Section 4.2 to correct for the selection bias introduced by the current allocation policy.

The estimated survival model can be used for various purposes. One purpose is testing hypotheses regarding effects of covariates on survival and comparing survival curves of subgroups. Another is characterizing patient lifetimes. A third purpose is predicting PTLY for new waitlist patients. We discuss these purposes further in Section 7.

6. Applying SimBa to the U.S. Kidney Waitlist

In the following, we estimate pre-transplant survival models based on the current waitlist of registrations and transplants of kidney and simultaneous kidney-pancreas in the United States. We compare a weighted AFT model based on SimBa propensity scores to two existing approaches: an ordinary AFT model applied only to candidates' data (the CCA approach), and a weighted AFT model that uses logistic regression propensity scores as weights. We start by describing the data, then discuss the computation of propensity scores, and finally present the different survival models.

6.1 Data

We consider the dataset of waiting list registrations and transplants of kidney and simultaneous kidney-pancreas[2] that have been listed or performed in the U.S. and reported to the OPTN between October 1, 1987 and October 1, 2010. The dataset includes records on both deceased and living-donor transplants. The data were exclusively provided by UNOS[3].

Preliminary analysis of the data exhibits a rapid increase in patients' lifetime over the last 30 years, possibly due to changes in recent medical and dialysis treatments. Rather than incorporate these changes into our model,

[2] Simultaneous transplantation of a kidney and pancreas is performed for those who have kidney failure as a complication of insulin-dependent diabetes mellitus (also called Type I diabetes).

[3] The data reported here have been supplied by the United Network for Organ Sharing as the contractor for the Organ Procurement and Transplantation Network. The interpretation and reporting of these data are the responsibility of the author(s) and in no way should be seen as an official policy of or interpretation by the OPTN or the U.S. Government

as suggested in Mauger et al (1995), we consider a subset of the patients who *joined* the waiting list after January 2000. Truncating the data according to arrival time, as opposed to waiting list status on that year, ensures that patients' arrivals are approximately uniformly distributed[4] over the studied interval: [2000, 2010]. For computational reasons, we apply the policies and evaluate them for a single geographic region. We randomly chose region #2 that contains the following states: Pennsylvania (PA), New Jersey (NJ), West Virginia (WV), Maryland (MD), Delaware (DE), and Washington DC (DC). Our studied subset includes over 29,000 patients, among them 10,400 (approximately 36%) received transplants and about 4,300 (approximately 15%) died while waiting for transplant.

6.2 Computing Propensity Scores

To compute the propensity scores from an ordinary logistic regression model, we fit a model to the entire waitlist such that Y=transplant/no-transplant (recipient/candidate) and the covariates are the patient's health condition at the time of joining the waitlist and his/her location. The propensity score, or estimated probability of transplant, for patient i, given covariates $\mathbf{z}_i$, is

$$PS_i(\hat{\theta}) = \frac{\exp(\hat{\theta}' z_i)}{1 + \exp(\hat{\theta}' z_i)}, \tag{4}$$

where $\hat{\boldsymbol{\theta}}$ is the estimated parameter vector and $\mathbf{z}_i$ is the corresponding vector of covariates.

The estimated model using our data is provided in Table 1. The distribution of the resulting propensity scores is shown in Figure 4. The top left panel displays the scores for recipients and bottom left for candidates. Ideally, scores should be low for candidates and high for recipients. The figure shows that the logistic model captures candidates reasonably well (the histogram exhibits a right-tailed distribution), but does not capture recipients well (the histogram is bell shaped).

Next, we use SimBa to generate propensity scores, as described in Section 4. We generate $M = 100$ donor datasets, each replicated $N = 5$ times, for a total of $M \times N = 500$ runs.

The distribution of the propensity scores is shown in Figure 4. The top right panel corresponds to recipients' scores and the bottom right to candidates. Overall, SimBa captures both candidates' and recipients' scores quite well, polarizing the scores of the two classes, such that most recipients receive near-one probabilities of transplant and candidates near-zero probabilities, with a few exceptions.These exceptions give an idea of the KPSAM simulator's inaccuracy. Let us examine the exceptions and their magnitude. For candidates, the great majority were assigned zero or near-zero scores. Approximately 0.7% of candidates (∼130 candidates) received a score of 1, but in reality did not

[4] Under the assumption of Poisson arrival times, as evidenced from the data.

Table 1 Estimated Logistic Regression Parameters from Selection Probability Model

Covariate (Z_k)	$\hat{\theta}_k$	$SE(\hat{\theta}_k)$	$p-value$
State: Delaware	0.36	0.12	0.00
State: Maryland	0.63	0.10	0.00
State: New Jersey	0.37	0.10	0.00
State: Pennsylvania	0.66	0.09	0.00
State: West Virginia	0.99	0.12	0.00
Human Leukocyte Antigen (HLA)	0.31	0.08	0.00
Diabetes	-1.34	0.20	0.00
Simultaneous kidney-pancreas (KP)	0.57	0.09	0.00
Dialysis	0.23	0.06	0.00
Previous transplant (yes/no)	-0.58	0.05	0.00
Albumin	-0.24	0.04	0.00
Panel Reactive Body (PRA)>80	1.72	0.16	0.00
Diagnosis unknown	0.21	0.07	0.00
Polycystic kidneys	0.52	0.28	0.06
Male	-0.06	0.04	0.19
ABO type AB	0.42	0.07	0.00
ABO type B	-0.50	0.04	0.00
ABO type O	-0.43	0.03	0.00
Functional status: minor disability	-0.30	0.11	0.01
Functional status: some disability	-0.06	0.08	0.45
Age	-0.01	0.00	0.00
No Antigens	0.15	0.04	0.00
African American	-0.30	0.21	0.15
Diabetes x Diagnosis unknown	0.39	0.09	0.00
Diabetes x Age	0.02	0.00	0.00
Diabetes x Male	0.17	0.06	0.00
Diabetes x African American	0.18	0.06	0.00
Dialysis x African American	-0.14	0.07	0.04
Dialysis x Hospitalization History	-0.11	0.02	0.00
PRA>80 x Hospitalization History	-0.37	0.09	0.00
Male x African American	0.27	0.06	0.00
Previous transplant x Diagnosis unknown	0.19	0.13	0.14
Some disability x Hospitalization History	0.09	0.03	0.00
Albumin x Hospitalization History	0.14	0.02	0.00
Albumin x African American	0.09	0.05	0.08
KP x PRA>80	-1.44	0.43	0.00
KP x Diagnosis unknown	-0.97	0.27	0.00
KP x Previous transplant	-0.97	0.31	0.00
KP x Age	-0.45	0.13	0.00
HLA x Diabetes	-0.26	0.12	0.03
Diagnosis unknown x Hospitalization History	-0.13	0.03	0.00
No Antigens x African American	-0.12	0.06	0.06

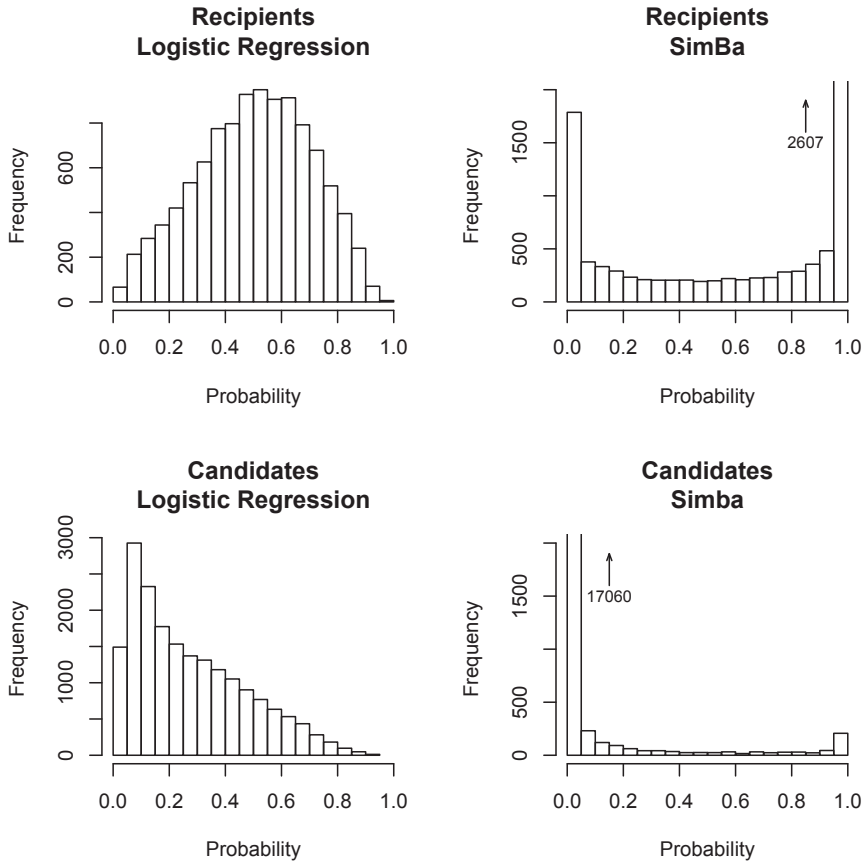

Fig. 4 Distribution of propensity scores of recipients and candidates.

receive transplant. For recipients, while most received a near-one score, about 10% were assigned a zero score (∼1200 patients) but in reality they did receive a kidney. The two types of discrepancies imply that some allocation considerations are not captured by the KPSAM simulator. We discuss this point further in Section 7.

We evaluate and compare the propensity scores generated by the logistic model and SimBa by plotting their lift charts in Figure 5. The lift charts clearly show that SimBa outperforms the logistic regression in separating recipients from candidates. The corresponding c-statistics ("area under the ROC curve"), which measure the likelihood that a randomly selected recipient has a higher score than a randomly selected candidate, are 0.78 for the logistic model, and 0.92 for SimBa. While both exceed the random 0.5, SimBa significantly outperforms the logistic regression in terms of lift.

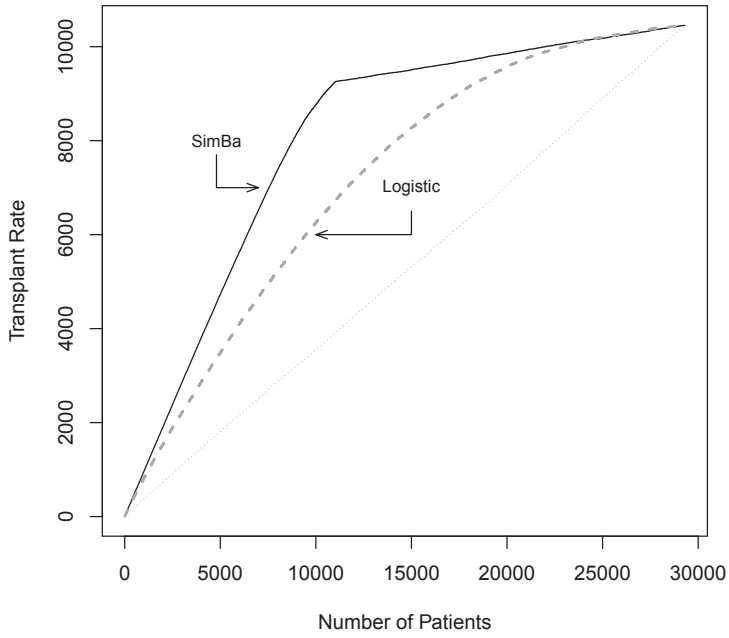

Fig. 5 Lift chart of the logistic and SimBa models

Weights for candidates are then generated using the propensity scores, for both the logistic regression and SimBa, using the binning procedure described in Section 4.3. In particular, we use 100 equal-size subgroups.

6.3 Modeling Pre-Transplant Survival Rates

We estimate the survival rates of waitlist patients' (candidates and recipients) using a weighted Weibull AFT model that uses SimBa-based weights . We compare the survival model with an ordinary Weibull AFT model applied only to candidates' data (the CCA approach), and a weighted Weibull AFT model that uses logistic regression-based weights. Based on the model in Wolfe et al (2008), the covariates in each of the survival models include age, disease history (such as reason for listing: kidney alone or simultaneous kidney-pancreas (KP), years with dialysis, and previous transplant) primary diagnosis at time of listing (such as diabetes, dialysis, and polycystic kidneys), and various interactions between these covariates. The estimated models are given in Table 2, and compared graphically in Appendix B, Figure 14. While the differences between coefficients across the three models are mostly insignificant, for some covariates there is a significant difference. For example, the coefficient for AGE

is significantly larger in magnitude for SimBa, indicating a stronger correlation between survival and age. For Previous Transplant, only the logistic model has a statistically significant coefficient at 5% significance level.

To compare the models, we plot the baseline survival curves in Figure 6 and PTLY distribution in Figure 7. The results reveal contradicting pictures of the relationship between selection-to-transplant probability and survival rate. Let us first consider two main results:

1. Unsurprisingly, the logistic regression approach is more pessimistic than the CCA approach in terms of patients' survival rates and lifetimes (due to CCA modeling only candidates' data). In fact, according to the logistic regression model, the great majority of patients do not live longer than 20 years. The average PTLY under this model is approximately 9 years, compared to nearly 11 years under CCA approach.
2. The SimBa approach is more optimistic than both alternatives. The average PTLY under SimBa is approximately 13.5 years, 25% longer than the CCA average, and 51% longer than the logistic average.

Recall that the logistic model includes both candidates and recipients. Because the expected PTLY under this model is shorter than under the candidates-only CCA model, it necessarily indicates that recipients' survival rates are lower. In other words, the logistic model indicates that transplant probability *decreases* in expected PTLY. In contrast, the SimBa model indicates that transplant probability *increases* in expected PTLY, implying also that the CCA model *under-estimates* the survival rate of waitlist patients.

The correctness of the SimBa approach conclusion over the logistic regression approach conclusion can be explained systematically and theoretically. We also show that it is supported empirically. Systematically, two factors contribute to a negative correlation between transplant probability and survival rate. The first factor is that patients with a severe health condition are typically not considered by the current policy to be appropriate candidates for most available organs, and consequently their transplant probability is low. The second factor is the high rate of kidney refusal decisions; it is well known that the current allocation system poorly matches candidates and kidneys (see OPTN/UNOS (2008)). This results in a high patient refusal rate (45%, according to Zenios (2004)), because patients and their physicians anticipate transplant failure or degradation in the patient's post-transplant health condition or quality of life. Hence, patients with lower PTLY tend to wait longer for suitable organs.

Theoretically, the assumption of strong ignorability in the logistic regression approach means that by-design the propensity scores, and hence the resulting survival model, cannot correctly capture the relationship between the transplant probability and the actual outcome. The logistic approach models each of these components separately: the transplant probability is modeled in the first step (propensity scores computed via logistic regression, as described in Section 6.2), where the anticipated outcome is ignored; the outcome is then modeled by the AFT model. In contrast, SimBa considers the indirect effects

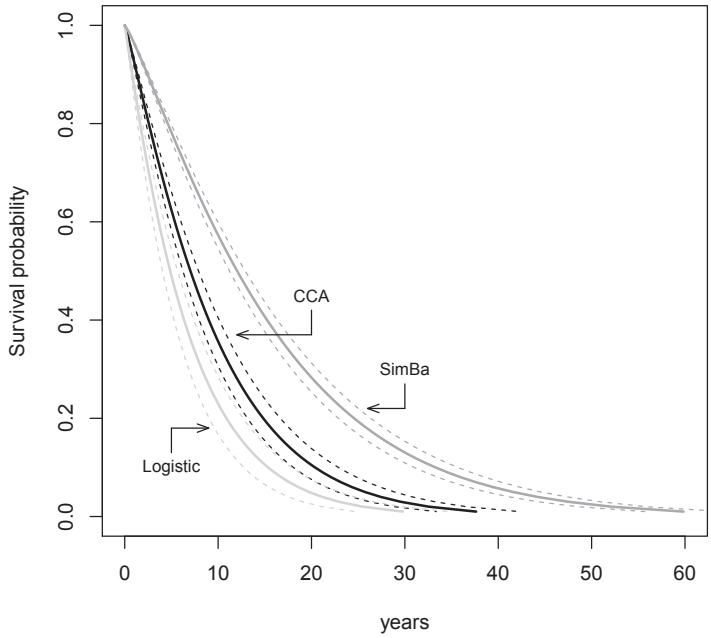

Fig. 6 Estimated baseline survival rates (with 95% confidence intervals), by model.

of the outcome on transplant probability, and therefore captures this positive relationship between transplant probability and PTLY.

Empirically, we find evidence in the waitlist dataset showing that patients with more severe health conditions (e.g., diabetics vs. non-diabetics) have a higher mortality rate, but a lower transplant rate. For example, among diabetic candidates 26% died while waiting for transplant compared to a death rate of 20% for non-diabetic patients. Yet, the transplant rate among diabetic candidates is less than 30% as opposed to 40% among non-diabetics. Similar patterns are found for other health indicators such as age, time on dialysis, and the need for simultaneous kidney and pancreas transplant, which is more severe than a need for only a kidney. These patterns are shown in Figures 8 and 9. Figure 8 compares high-risk ("H") and low-risk ("L") groups in terms of death rate and transplant rate for selected health covariates that exhibited such differences. Note that this pattern was not observed for covariates Dialysis, Unknown diagnosis, and PRA. Figure 9 compares death rate and transplant rate for two numerical covariates binned into subgroups: age and dialysis years. Both figures show that patients with more severe health conditions have a higher mortality rate but lower transplant rate compared to "healthier" patients.

Table 2 Estimated Parameters (and standard errors) from Weibull AFT Models

Covariate (Z_k)	CCA: $\hat{\gamma}_k$ ($SE(\hat{\gamma}_k)$)	Logistic: $\hat{\gamma}_k$ ($SE(\hat{\gamma}_k)$)	SimBa: $\hat{\gamma}_k$ ($SE(\hat{\gamma}_k)$)
Diabetes	-0.70 (0.18)	-0.53 (0.19)	-0.96 (0.20)
Simultaneous kidney-pancreas	-1.38 (0.23)	-1.26 (0.20)	-1.69 (0.26)
Dialysis	-0.45 (0.03)	-0.45 (0.03)	-0.50 (0.03)
Previous transplant (yes/no)	-0.06 (0.03)	**-0.09 (0.03)**	-0.01 (0.04)
Time on dialysis	0.02 (0.01)	0.03 (0.01)	0.00 (0.01)
Age	**-0.02 (0.00)**	**-0.02 (0.00)**	**-0.03 (0.00)**
Diagnosis unknown	-0.13 (0.03)	-0.21 (0.03)	-0.13 (0.03)
Polycystic kidneys	0.31 (0.07)	0.20 (0.06)	0.40 (0.07)
Body Mass Index (BMI)	0.02 (0.00)	0.02 (0.00)	0.02 (0.00)
Albumin	0.34 (0.03)	0.35 (0.03)	0.34 (0.03)
Number of A antigens	-0.14 (0.05)	-0.17 (0.05)	-0.10 (0.05)
Number of B antigens	-0.16 (0.04)	-0.13 (0.04)	-0.17 (0.05)
Number of DR antigens	-0.22 (0.05)	-0.31 (0.05)	-0.26 (0.05)
No Antigens	-0.22 (0.05)	-0.26 (0.05)	-0.24 (0.06)
Panel Reactive Body (PRA)	-0.01 (0.00)	-0.01 (0.00)	-0.01 (0.00)
KP x Diabetes	1.97 (0.52)	2.51 (0.50)	2.61 (0.60)
Diabetes x Age	0.01 (0.00)	0.01 (0.00)	0.01 (0.00)
Diabetes x albumin	-0.07 (0.04)	-0.07 (0.04)	-0.04 (0.05)
Diabetes x Albumin x KP	-0.19 (0.10)	-0.46 (0.10)	-0.29 (0.12)
Shape	1.13	1.04	1.18
Scale	9.73	6.91	16.42

6.4 Summary of Findings

We apply the proposed SimBa approach, which is a simulator-based method for generating propensity scores, to adjust for selection bias in modeling pre-transplant survival rates using the 2000-2010 U.S. kidney transplant waiting list. We compare the weighted survival model based on SimBa propensity scores to two existing approaches: an ordinary survival model based only on candidates' data (CCA approach) and a weighted survival model based on logistic regression propensity scores. The main findings are as follows:

1. SimBa generates a more accurate set of propensity scores compared to the logistic regression model, better capturing the actual transplant selections. SimBa's lift outperforms the lift of the logistic regression, with a c-statistic of 0.92, compared to the logistic model's 0.78 (an improvement of 18%).
2. The SimBa-based survival model is more optimistic than both CCA and the logistic models in terms of survival rates. This implies that the CCA survival model under-estimates patients' survival rates. It also means that the SimBa model captures a positive correlation between transplant prob-

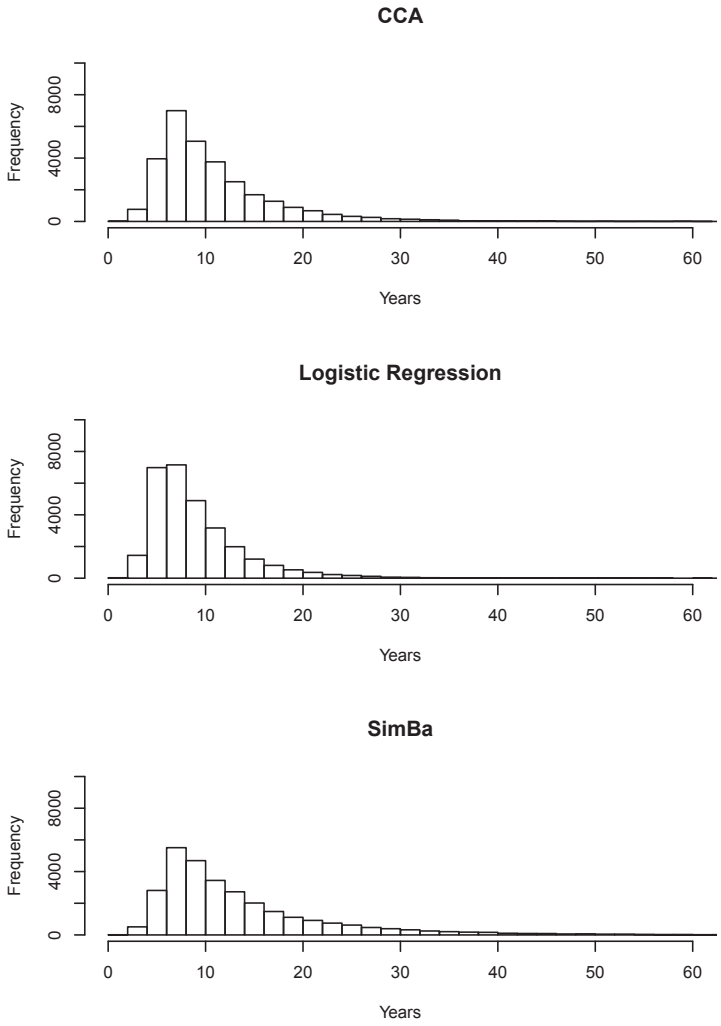

Fig. 7 Estimated baseline lifetime distribution, by model.

ability and pre-treatment survival rate. That is, a longer expected lifetime without transplant is positively correlated with transplant probability.

3. Relying on an assumption of strong ignorability, the logistic model in fact increases selection bias instead of reducing it, thereby resulting in overly pessimistic survival rates compared to both the CCA and SimBa approaches.

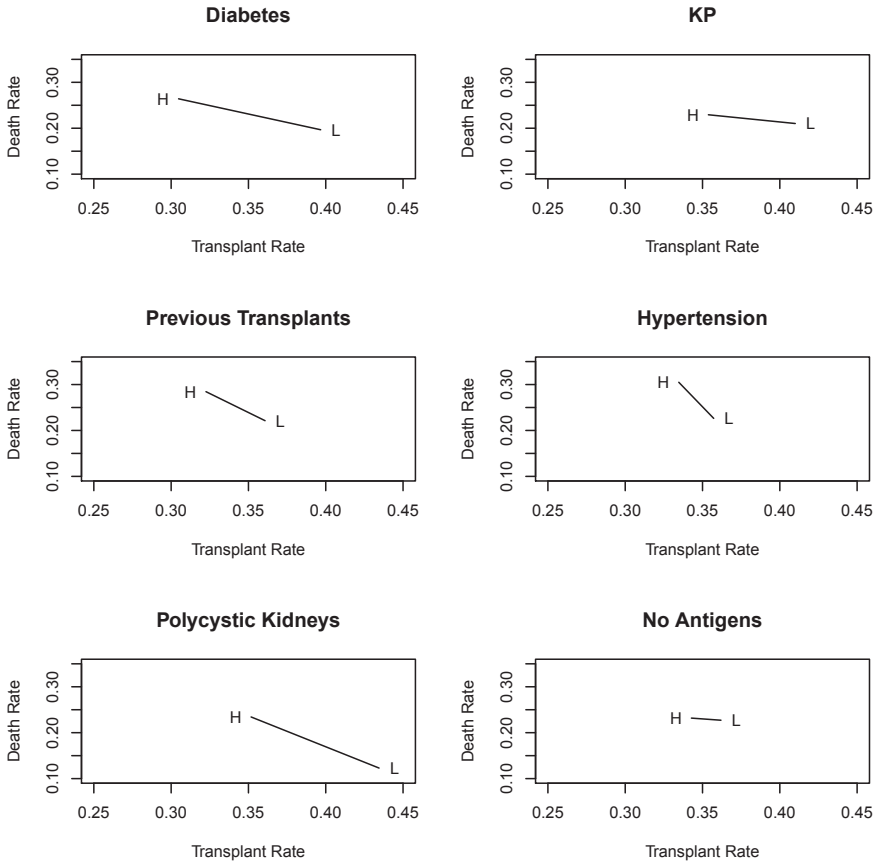

Fig. 8 Transplant and death rates of High (H) and Low (L) risk candidates (binary covariates).

6.5 Sensitivity Analysis

The analysis presented in this section raises some additional questions regarding the robustness of SimBa, its sensitivity to data selection, and how it compares to other nonparametric methods. In this section we address these questions. Our analysis examines the robustness of the SimBa survival curves to three factors: (1) the choice of study duration, (2) aggregation of study regions, (3) using non parametric methods to compute propensity scores. Finally, we compare the SimBa survival curves to those based on a data imputation-based model.

The first set of analyses examines the effect of medical technology changes on survival rates. For that, we re-estimate the survival curves under different durations within the same region. In particular, we look at the last three years, the last five years, and the last ten years. We examine the duration

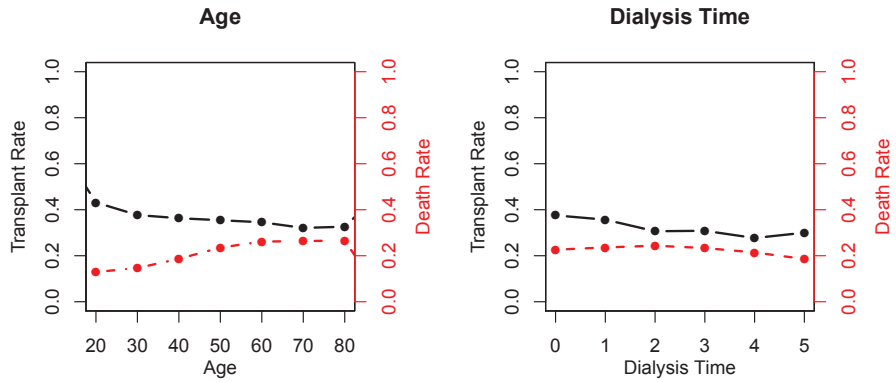

Fig. 9 Transplant and death rates of candidates (binned numerical covariates).

effect only on the CCA model for several reasons. First, CCA models the survival of candidates (who have not received a kidney). These patients are expected to be the most affected by changes in medicines prescribed for dialysis patients. Second, the KPSAM simulator does not model changes in medication and assumes that the same medication is used throughout the study period. Third, the computation time of KPSAM can be prohibitive: each run for a single region takes approximately 4 hours, and we execute 100 runs. Hence, for a particular region, our analysis requires 400 hours of computation. But most importantly, the average waiting time on the waitlist in our dataset is approximately 4.5 years. Hence, modeling selection-to-transplant for a short period, such as 3 or 5 years, is practically impossible with SimBa or logistic regression.

We find that the survival curves are practically identical in all cases. This can be seen in Figure 10, showing the three curves with 95% confidence bands. We conclude that even if medical technology does change over time, survival curve estimation is less sensitive to such changes in terms of producing similar survival curves.

The second set of analyses compares survival rates in different regions, to see whether all regions can be modeled within a single model. Using a five-year period, we compared Region 2 alone (States: PA, NJ, WV, MD, DE, DC) to all regions. We also compared Region 2 with a combination of Regions 2 and Region 9 (NY and Western VT). These two regions were chosen because the literature shows that they are statistically similar in terms of selection to transplant, health condition, disease and mortality rates (Mathur et al, 2010; Garonzik-Wang et al, 2012). The results using CCA are shown in Figure 11. We see that the survival curves for the three choices of region combinations differ, especially in longer survival durations. Even combining two regions that are supposed to be statistically similar leads to different survival curves.

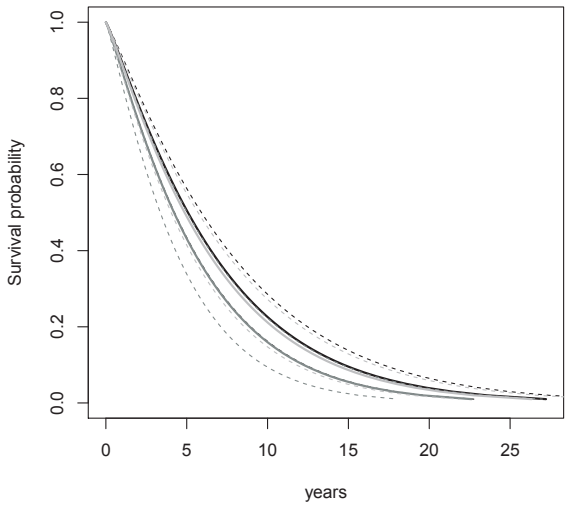

Fig. 10 Survival curves based on 3, 5 and 10 year span of data from region 2. (black: 10 years, dark grey: 5 years, light grey: 3 years) .

For computational reasons we do not examine the results using SimBa. However, because even CCA results in different curves for different regions, the same will be the case with the propensity-score-based models, which are essentially weighted versions of the CCA survival model. Hence, we conclude that modeling should be performed separately for different regions.

Third, we consider replacing the logistic regression model with a random forest for computing propensity scores (Lee et al, 2010). The advantage of random forests over logistic regression is that they are nonparametric and data-driven. With a large sample, such an approach can capture unexpected relationships between the covariates and outcome variable. In our case, we find that the results of the two approaches are very similar, with a minuscule advantage for the random forest in terms of lift (c-statistic of 0.79 for random forest vs. 0.78 for the logistic model; see Figure 12). The reason for the similarity is that trees, like the logistic model, are only based on the waitlist data and not on the KPSAM simulator results. Hence, they too suffer from violating the strong ignorability assumption.

Lastly, we compare the performance of SimBa to that of data imputation models for estimating survival rates. In particular, we use a random forest to impute the missing PTLY and death incidences for recipients and re-run the AFT model on the completed set of patients. We use the method from Yahav and Shmueli (2010), which consists of two steps: first, for recipients we predict whether they would have died, had they not received the kidney. This is done using a classification random forest. Then, for those predicted to have died, we

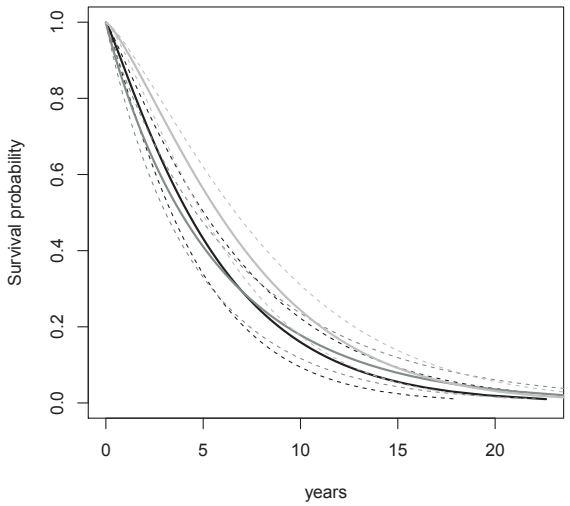

Fig. 11 Survival curves under CCA, using data from the last five years and different sets of regions (black: Region 2, dark grey: Regions 2 and 9, grey: all regions).

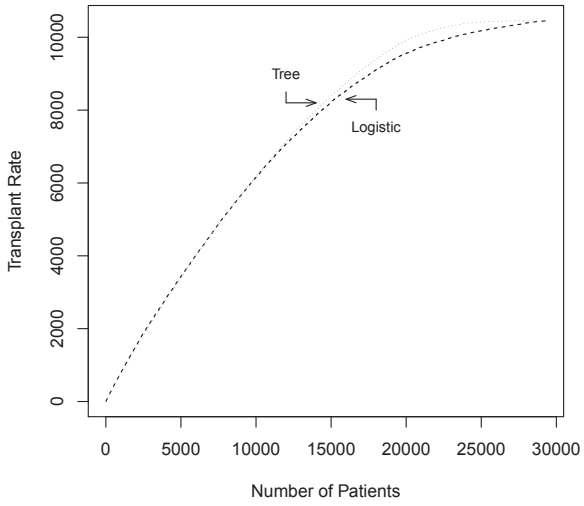

Fig. 12 ROC curve of propensity scores under logistic model compared with random forest.

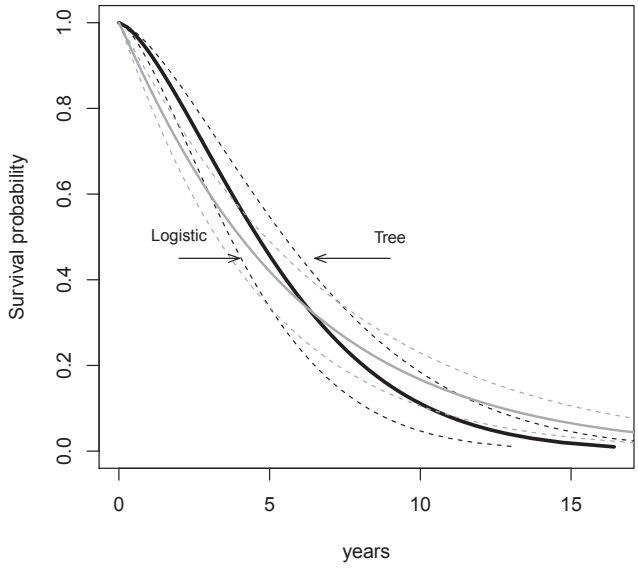

Fig. 13 Survival curves and 95% confidence bands under random forest (black) and logistic (grey) models.

estimate their PTLY using a regression random forest. Our analysis shows that the performance is similar to that of the logistic regression (see Figure 13). The reason is that in order to predict each of the model steps (that is, death incidences and PTLY), we must assume that the data is complete (CCA), or unbiased. We predict death incidences based on the candidate pool (those who have not received transplant), assuming they also represent the recipients. In other words, we assume that the selection to transplant does not depend on PTLY, which we showed to be untrue.

7. Discussion and Implications

The main question raised in this paper is whether allocation outcomes matter when estimating pre-transplant survival rates of kidney-transplant patients. The results of our simulation approach clearly show that the answer to this question is *yes*. Moreover, disregarding the outcome and relying on the assumptions of strong ignorability while ignoring censoring due to the study period introduces an even greater selection bias into survival rates than a naïve, complete-case model.

Our conclusion from this analysis is the insufficiency of waitlist data alone for estimating pre-transplant survival rates. Hence, it is necessary to integrate

it with data on outcomes and other sources of uncertainty by using the KPSAM simulator. This insufficiency of the waitlist data is related to the notion of "information quality", which is "the potential of a dataset to achieve a specific (scientific or practical) goal using a given empirical analysis method" (Kenett and Shmueli, 2011). Clearly, the information quality of the waitlist data for estimating pre-transplant survival rates using classic propensity scores is low, and can be drastically increased by including outcome information via the KPSAM simulator.

Basing our model on the KPSAM simulator, has some limitations that are derived from the simulator assumptions. In particular, two models underlie the KPSAM simulator:

Kidney acceptance model (whether a patient accepts a kidney offered to him/her) we do not know how this model is computed and whether it depends on survival rates. In practice, there are personal considerations that cannot be modeled, which can lead to outliers, as described in Section 6.2

Patient pre-transplant survival model (for survival without a transplant) the survival model used in KPSAM is a Cox regression model. According to our results, we find that using the simulator to predict recipients lifetime yields shorter lifetimes than expected. However, since the KPSAM assignment model is fairly accurate (see Figures 4-5), the estimate of candidates lifetime under KPSAM does not significantly affect the SimBa estimates (recall that we only use KPSAM for computing scores and not for estimating lifetimes).

While the KPSAM simulator is not perfect and ignores some factors affecting kidney allocation, it offers a significant improvement over using waitlist data alone, as well as offers insights into the current Priority Points allocation policy. The 10% of misclassified recipients that we found in our dataset (which appear to be a random sample of recipients) means that the SimBa survival curve is a lower bound for a similar curve that would be based on a perfect simulator.

Additional uses of pre-transplant survival models are (1) inference regarding the effects of different health covariates on survival, (2) characterization of patient lifetimes, and (3) predictions of individual patients' survival or PTLY. With respect to inference, in our data (the 2000-2010 US waitlist) the three estimated survival models are very similar for most covariates, except for the significance of Previous Transplant and the magnitude of Age (see Table 2). Hence, bias can be a concern. In terms of PTLY distribution, our example shows that estimated lifetimes of patients under the different models can differ substantially (see Figure 7).

Predicting the pre-transplant survival rates of new patients is an important component of new guidelines for allocation. In 2004, the Kidney Allocation Review Subcommittee (KARS) was established with the goal of designing an allocation policy that maximizes the tradeoff between *equity* in access to transplantation and *efficiency*; that is, maximizing the aggregate health of the transplant candidate pool (Votruba, 2001). In 2008, the committee proposed

four concepts that would together combine to determine a candidate's Kidney Allocation Score (KAS). One of the key concepts is *Life Years From Transplant (LYFT)*, which is the difference between the Post-Treatment Life Years and PTLY. The KAS equation assigns scores based on LYFT and three other parameters: (1) Donor Profile Index (DPI, ranges between 0 and 1), which measures the organ quality based on clinical information, (2) Calculated Panel Reactive Antibody (CPRA), which is the likelihood that the recipient and donor are incompatible, and (3) time on dialysis time (DT). LYFT, DPI and CPRA together measure the *efficiency* of the allocation policy, while DT is a measure of *equity*.

To evaluate the effect of under-estimating PTLY on the new allocation policy We compared PTLY values of candidates under CCA and SimBa and found that the correlation is nearly 0.98. This implies that the ordering of PTLY values (and consecutively, LYFT) for a set of patients will remain the same under CCA and SimBa. However, the order of transplant probability *significantly changes*. Moreover, the total weight of the LYFT component (the policys efficiency) in the KAS equation is in practice higher than it should be, due to the current under-estimation of PTLY values. Since there is a tradeoff between efficiency and fairness (measured by DT in the KAS equation), giving too much weight to efficiency leads to a decreased fairness weight.

In the context of predicting PTLY values for new patients, while it is easy to generate predictions, it is more challenging to assess the predictive accuracy of the survival models. Evaluating the predictive power of a model would ideally be based on the predictive performance on a holdout set. As is customary in predictive modeling, we would ideally fit the survival models to a training set and then use this model to predict PTLY values of patients in a holdout set. Unfortunately, the waitlist dataset consists mostly of unlabeled data, where we do not know the actual PTLY values, neither for recipients nor for live candidates. The only patients for whom we have complete PTLY values are deceased candidates. This group consists of approximately 13.5% of our dataset[5].

In our analyses, we included all information that was available in the waitlist data for generating simulated outcomes and for estimating survival rates. We included the few sporadic health updates (where different patients had different updates at different times) by using the KPSAM simulator. Potential improvement to pre-transplant survival models can be obtained if waitlist data contain frequently updated health data in a structured form. Such updates can then be integrated into the model via time-varying covariates in the imputation model (as proposed in Wolfe (2007); Sela and Simonoff (2011)), the propensity model, or even directly in the survival model.

[5] We examined predictive accuracy for the deceased group based on separate training and holdout samples. While all three survival models generated mostly near-zero prediction errors, SimBa was much better at predicting PTLY values of deceased candidates who joined the waitlist late in the study and died early. This result emphasizes the importance of including the arrival time (rather than only the PTLY) in the model

One main limitation of the SimBa model is its high computation time: each run for a single region takes approximately 4 hours, and we run 100 runs in each case. Hence, for a particular region, our analysis requires 400 hours of computation. Note that such computations are needed only once per region for generating the propensity scores and estimating survival rates. However, periodic updates of the SimBa model, additional sensitivity tests, and other model changes require additional extensive simulation runs. This limitation, which applies only to future uses of SimBa, will become less prohibitive as computation power increases. Regardless of computational issues, our current study revealed the important point that allocation outcomes matter when estimating pre-transplant survival rates of kidney-transplant patients. This result in itself highlights a feature of the current allocation policy, questions the validity of current methodology for estimating pre-treatment survival rates and the need for methods that properly account for the outcome effect.

An important question that our results raise is the robustness of alternative allocation policies that incorporate PTLY into the allocation decisions to ignoring strong ignorability and time-changing covariates. Would a biased survival model significantly affect allocation decisions of a newly designed policy? The impact of the survival model and selection bias on allocation policies is an open question that requires further research.

References

Bavaria J, Appoo J, Makaroun M, Verter J, Yu Z, Mitchell R, Gore T (2007) Endovascular stent grafting versus open surgical repair of descending thoracic aortic aneurysms in low-risk patients: a multicenter comparative trial. The Journal of Thoracic and Cardiovascular Surgery 133(2):369–377

Binder D (1992) Fitting Cox's proportional hazards models from survey data. Biometrika 79(1):139–147

Cho K, Himmelfarb J, Paganini E, Ikizler T, Soroko S, Mehta R, Chertow G (2006) Survival by dialysis modality in critically ill patients with acute kidney injury. Journal of the American Society of Nephrology 17(11):3132

D'Agostino RBJ (2007) Propensity scores in cardiovascular research. Circulation 115(17):23–40

D'Agostino RJ (1998) Tutorial in Biostatistics: Propensity Score methods for bias reduction in the comparison of a treatment to a non randomized control group. Statistics in Medicine 17:2265–2281

Demissie S, LaValley M, Horton N, Glynn R, Cupples L (2003) Bias due to missing exposure data using complete-case analysis in the proportional hazards regression model. Statistics in Medicine 22(4):545–557

Efron B, Tibshirani R (1993) An introduction to the bootstrap, vol 57. Chapman & Hall/CRC

Engoren M, Habib R, Zacharias A, Schwann T, Riordan C, Durham S (2002) Effect of blood transfusion on long-term survival after cardiac operation. The Annals of thoracic surgery 74(4):1180–1186

Field C, Smith B (1994) Robust estimation: a weighted maximum likelihood approach. International Statistical Review 62(3):405–424

Garonzik-Wang J, James N, Weatherspoon K, Deshpande N, Berger J, Hall E, Montgomery R, Segev D (2012) The aggressive phenotype: Center-level patterns in the utilization of suboptimal kidneys. American Journal of Transplantation

Henry GT (1990) Practical Sampling. SAGE Publications

Kenett RS, Shmueli G (2011) On Information Quality *Working Paper RHS 06-100*. Robert H Smith School of Business, University of Maryland

Kleinbaum DG, Klein M (1995) Survival analysis: a self-learning text. Springer

KPSAM (2009) Kidney-Pancreas Simulated Allocation Model. Arbor Research Collaborative for Health, Scientific Registry of Transplant Recipients, 4th edn

Lambert P, Collett D, Kimber A, Johnson R (2004) Parametric accelerated failure time models with random effects and an application to kidney transplant survival. Statistics in medicine 23(20):3177–3192

Lee B, Lessler J, Stuart E (2010) Improving propensity score weighting using machine learning. Statistics in medicine 29(3):337–346

Lin D (2000) On fitting Cox's proportional hazards models to survey data. Biometrika 87(1):37–47

Mathur A, Ashby V, Sands R, Wolfe R (2010) Geographic variation in end-stage renal disease incidence and access to deceased donor kidney transplantation. American Journal of Transplantation 10(4p2):1069–1080

Mauger E, Wolfe R, Port F (1995) Transient effects in the Cox proportional hazards regression model. Statistics in Medicine 14(14):1553–1565

OPTN/UNOS (2008) Kidney Allocation Concepts, Request for Information. The Kidney Transplantation Comittee

Pan Q, Schaubel D (2008) Proportional hazards models based on biased samples and estimated selection probabilities. Canadian Journal of Statistics 36(1):111–127

Pan Q, Schaubel D (2009) Evaluating bias correction in weighted proportional hazards regression. Lifetime Data Analysis 15(1):120–146

Polkinghorne K, McDonald S, Atkins R, Kerr P (2004) Vascular access and all-cause mortality: a propensity score analysis. Journal of the American Society of Nephrology 15(2):477–486

Rosenbaum PR, Rubin DB (1983) The central role of the propensity score in observational studies for causal effects. Biometrika 70(1):41–55

Rosenbaum PR, Rubin DB (1984) Reducing Bias in observational studies using subclassification on the propensity score. Journal of the American Statistical Association 79(387):516–524

Schaubel D, Guidinger M, Biggins S, Kalbfleisch J, Pomfret E, Sharma P, Merion R (2009) Survival benefit-based deceased-donor liver allocation. American Journal of Transplantation 9(4p2):970–981

Sela R, Simonoff J (2011) Re-em trees: a data mining approach for longitudinal and clustered data. Machine Learning pp 1–39

Shyur H, Elsayed E, Luxhøj J (1999) A general hazard regression model for accelerated life testing. Annals of Operations Research (Special Issue on Reliability and Maintenance in Production Control) 91:263–280

SRTR (2007a) Methods for Discounting Median Lifetimes. *Working paper*

SRTR (2007b) Predicting the Life Years From Transplant (LYFT): Choosing a Metric. *Working paper*

Stürmer T, Schneeweiss S, Brookhart M, Rothman K, Avorn J, Glynn R (2005) Analytic strategies to adjust confounding using exposure propensity scores and disease risk scores: nonsteroidal antiinflammatory drugs and short-term mortality in the elderly. American Journal of Epidemiology 161(9):891–898

UNOS UNoOS (2011) Allocation of deceased kidneys (3.5)

Votruba M (2001) Efficiency-equity tradeoffs in the allocation of cadaveric kidneys *Working Paper*. Princeton University

Wolfe R (2007) Avoiding statistical bias in predicting the life years from transplant (LYFT). *Working paper*. SRTR Working Paper

Wolfe R, McCullougha K, Schaubelb D, Kalbfleischb J, Murrayb S, Stegallc M, Leichtmanb A (2008) Calculating Life Years from Transplant (LYFT): Methods for Kidney and Kidney-Pancreas Candidates. American Journal of Transplantation 2008(8 part 2):997–1011

Yahav I, Shmueli G (2010) Predicting potential survival rates of kidney transplant candidates from databases with existing allocation policies

Zenios S (2004) Models for kidney allocation. Operations Research and Health Care: A Handbook of Methods and Applications pp 537–554

Appendix A: Description of Variables in U.S. Kidney Waitlist Dataset

Below is a list of variables used throughout the paper. We list their abbreviation and description.

ABO type AB (ABO_AB) Patient's blood group is AB
ABO type B (ABO_B) Patient's blood group is B
ABO type O (ABO_O) Patient's blood group is O
Age (AGE) Patient's age *upon arrival*
African American (AFRICAN) Patient's race is African American
Albumin (ALBUMIN) Patient albumin level. Low albumin levels reflect possibility of diseases in which the kidneys cannot prevent albumin from leaking from the blood into the urine and being lost
Body Mass Index (BMI) Patient's Body Mass Index (ratio of weight to square root of the hight). BMI provided a measure of a patient's overweight (BMI>25) or underweight (BMI<18.5)
Diabetes (DIAB) Indicates whether a patient is diabetic
Diagnosis unknown (NotSPECIFIED) Indicates whether a patient has no diagnosis
Dialysis (DIAL) Indicates whether a patient needs dialysis
Functional status: minor disability (MINOR_DIS) Patient can function with no assistance
Functional status: some disability (SOME_DIS) Patient can function with little assistance
Hospitalization History (HOSPITALIZATION) Number of previous hospitalizations
Hypertension (HYPERTENSION) Indicates whether a patient was diagnosed with malignant hypertension (a complication of hypertension characterized by very elevated blood pressure)
Human Leukocyte Antigen (HLA) Mean patient's antigen match with donors pool (ranges between [0,6])
Male (MALE) Patient's gender is male
No Antigens (ABDR) Indicates weather the patients has no antigens
Number of A antigens (A) Number of a patient's A antigens
Number of B antigens (B) Number of a patient's B antigens
Number of DR antigens (DR) Number of a patient's DR antigens
Panel Reactive Body (PRA) Patient's Panel Reactive Antibody (PRA) level (measure for sensitization level)
Polycystic kidneys (POLYCYSTIC) Indicates weather a patient was diagnosed with polycystic kidney syndrome (a genetic disorder that results in massive enlargement of the kidneys)
Previous transplant (PrevTRANS) Indicates whether a patient had previous transplants
Simultaneous kidney-pancreas (KP) Indicates whether a patient is waiting for simultaneous pancreas-kidney transplant
Time on dialysis (DT) Dialysis time in years *upon arrival*

Appendix B: Graphical Comparison Between the AFT Models

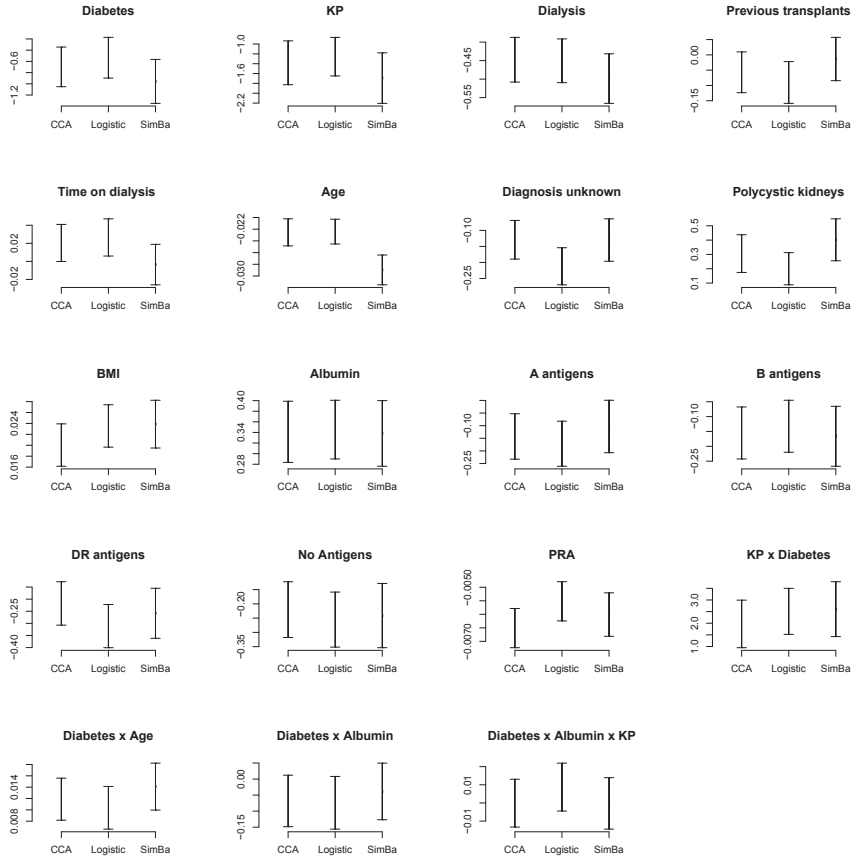

Fig. 14 Comparison of CCA, logistic, and SimBa AFT model coefficients. Each line corresponds to the parameter's 95% confidence interval.