A Comprehensive Data-Driven Characterization of Organ Transplantation

by

Diego Marconi Pinheiro Ferreira Silva

Master of Systems Engineering
University of Pernambuco
2013

Bachelor of Computer Engineering
University of Pernambuco
2011

A Dissertation submitted to Florida Institute of Technology in partial fulfillment of the requirements for the degree of

Doctorate of Philosophy in Computer Science

Melbourne, Florida
July, 2018
© Copyright 2018 Diego Marconi Pinheiro Ferreira Silva
All Rights Reserved

The author grants permission to make single copies.
We the undersigned committee hereby approve the attached Dissertation

A Comprehensive Data-Driven Characterization of Organ Transplantation by Diego Marconi Pinheiro Ferreira Silva

Ronaldo Menezes, Ph.D.
Professor, School of Computing
Committee Chair

Nezamoddin Nezamoddini-Kachouie, Ph.D.
Assistant Professor, Mathematical Sciences
Outside Committee Member

William Allen, Ph.D.
Associate Professor, School of Computing
Committee Member

Eraldo Ribeiro, Ph.D.
Associate Professor, School of Computing
Committee Member

Carmelo Bastos-Filho, Ph.D.
Associate Professor, University of Pernambuco, Brazil
Committee Member

Martin Cadeiras, M.D.
Assistant Professor, University of California, Los Angeles
Committee Member

Philip Bernhard, Ph.D.
Associate Professor and Director, School of Computing
ABSTRACT

Title: A Comprehensive Data-Driven Characterization of Organ Transplantation

Author: Diego Marconi Pinheiro Ferreira Silva

Major Advisor: Ronaldo Menezes, Ph.D.

Organ transplantation yearly saves thousands of lives worldwide; yet, 22 people still die each day only in the USA due to the ever-increasing imbalance between the supply and demand of organs. Currently, this organ-allocation gap is mainly tackled by optimally allocating organs based on major survivability factors (i.e., efficiency), by providing the population equal access to transplantation (i.e., equity), and by promoting population health literacy (i.e., awareness). Efficiency, equity, and awareness impact each other; yet, the state-of-the-art in organ transplantation still lacks the characterization of awareness, and the trade-off between these aspects are not fully-understood. Given the current availability of data and computational power, this dissertation proposes a comprehensive data-driven characterization of organ transplantation that accounts for efficiency, equity, and awareness by (i) integrating available data sets; (ii) proposing a sensor for population awareness using social media; and (iii) proposing a novel nonparametric probabilistic data-intensive framework. The dissertation demonstrates that the proposed characterization can uncover patterns of efficiency, equity, and awareness, and has the potential to characterize organ transplantation in a unified manner.
# Table of Contents

Abstract ........................................ iii

List of Figures ........................................ vi

List of Tables ........................................ xxxii

Acknowledgments ........................................ xxxv

1 Introduction ........................................ 1

2 Background ........................................ 9
    2.1 Organ Transplantation ............................ 9
    2.2 Data Science .................................... 17

3 Methodology ........................................ 20
    3.1 Curating and Integrating Data Sets ............. 21
    3.2 Creating a Sensor for Organ Transplantation Awareness using Social Media .................. 30
    3.3 A Nonparametric Probabilistic Data-Intensive Framework ............................. 36

4 On the Efficiency of Organ Transplantation ......... 47
    4.1 Efficiency and Distance .......................... 51
List of Figures

1.1 In organ transplantation, the aspects of efficiency, equity, and awareness are intertwined. Currently, organ transplantation lacks a characterization of population awareness. A comprehensive characterization of organ transplantation needs to take into account all these aspects in an unified manner. . . . . . . . . . . . . . . . . . . . . 3

1.2 Overview of proposed methodology to characterize the efficiency, equity, and awareness of organ transplantation. 1. A sensor for population awareness about organ transplantation issues using social media. 2. Curation and integration of organ transplantation and population demographics data sets. 3. A nonparametric probabilistic data-intensive framework. 4. Characterizing organ transplantation with regards to efficiency, equity, and awareness. . . 4

3.1 Patients added to the waiting list in the USA during the period of 1998 to 2012 for the six major solid organs (extracted from [60]). . 23

3.2 Transplants performed during the year in the USA during the period of 1998 to 2012 for the six major solid organs (extracted from [60]). 23
3.3 Correlation between racial/ethnic, education, and income. The ethnic/racial variables divide the population into the following 7 categories: White, Black, Hispanic, Indian, Asian, Pacific, and Other. The educational variables divide the population into the following 4 categories: Elementary, High, College, and Bachelor. Considering its income in the preceding 12 months, the income variables divide the population into the following 10 categories in dollars: less than 10k, from 10k to 15k, from 15k to 25k, from 25k to 35k, from 35k to 50k, from 50k to 75k, from 75k to 100k, from 100k to 150k, from 150k to 200k, and higher than 200k.

3.4 Donation registration data at the state-level for the years 2009, 2012, and 2016. The registration of each state is normalized by the population of each state.

3.5 Donation registration data at the level of zip code for the county of Los Angeles, California. This data is accumulated donation registrations until 2013.

3.6 The pipeline of organ-related collected tweets. (1) The text of each collected tweet needs necessarily to contain at least one work from the context set and at least one word from the subject set. (2) The self-reported location of each tweet is converted into a structured location using a geographic tool such as OpenStreetMap. (3) Tweets are filtered to those belonging to users from the US.
3.7 Organ-related tweets collected by our proposed sensor. (a) Distribution of the number of users over difference organs. (b) Distribution of the number of users and tweets over the number of organs mentioned. .................................................. 34

3.8 Proposed Nonparametric Probabilistic Data-Intensive Framework. The three components are (1) Data binning and Partitioning, (2) Combination and Estimation, and (3) Data Visualization. (1) The supports of $X$ and $Y$ are partitioned into $\mathcal{X}$ and $\mathcal{Y}$, respectively, using a specific types of data binning. (2) For each selected combination $(x, y)$ in $(\mathcal{X}, \mathcal{Y})$, the predefined nonparametric probabilistic model $\Pr(Y \leq y | X > x)$ is estimated by $\tilde{P}_{x,y}$ according to the Wilson Score as calculated in Equation 3.5 (see Figure 3.9). (3) After the numerical values of all probabilities are obtained, our data-intensive framework can be visualized using a data-visualization tool such as a heat map, where each color is associated with a specific probability value. .................................................. 38

3.9 Estimation employed in our proposed data-intensive framework. (1) Definition of a nonparametric probabilistic model for $\tilde{P}_{x,y}$. (2) Calculation of the confidence interval of the estimator. (3) Calculation of the Wilson Score as the estimator. (4) Calculation of the standard error of the estimator. ................................. 40
3.10 Wilson score estimator for three different estimates of population proportion ($\hat{p}$, dashed-black): 0.1 (left), 0.5 (center), and 0.9 (right). Each plot shows the Wilson estimator ($\tilde{p}$, green), its lower confidence interval ($\tilde{p} + z_{\alpha/2} \cdot \sigma_{\tilde{p}}$, red), and its upper confidence interval ($\tilde{p} - z_{\alpha/2} \cdot \sigma_{\tilde{p}}$, yellow) for sample size values from 1 to 30.

3.11 Deep examination of the association between $X$ and $Y$ using our proposed data-intensive framework. The three components are (1) Data binning and Partitioning, (2) Combination and Estimation, and (3) Data Visualization. (1) the supports of $X$ and $Y$ are partitioned into $X'$ and $Y'$, respectively, using a specific types of data binning. (2) For each selected combination $(x, y_2)$ in $(X', y_2)$, the predefined nonparametric probabilistic model $Pr(Y \leq y | X > x)$ is estimated by $\tilde{p}_{x,y_2}$ according to the Wilson Score as calculated in Equation 3.5 (see Figure 3.9). (3) After the numerical values of all probabilities are obtained, their confidence intervals can be visualized using a data-visualization tool such as a error bar, where each bar identifies the lower and upper values of a confidence interval.
3.12 Profiles of our proposed data-intensive framework. The three components are (1) Data binning and Partitioning, (2) Combination and Estimation, and (3) Data Visualization. (1) The supports of $X$ and $Y$ are partitioned into $\mathcal{X}$ and $\mathcal{Y}$, respectively, using a specific types of data binning. (2) For each selected combination $(x, y)$ in $(\mathcal{X}, \mathcal{Y}')$, the predefined nonparametric probabilistic model $\Pr(Y \leq y|X > x)$ is estimated by $\hat{p}_{x,y}$ according to the Wilson Score as calculated in Equation 3.5 (see Figure 3.9). (3) After the numerical values of all probabilities are obtained, the model profile for each $y$ can be visualized using a data-visualization tool such as multiple curves, where each curve is associated with a specific $y$.

3.13 Benchmarking of our data-intensive framework using a 10,000 data-points sampled from a Multivariate Gaussian $\mathcal{N}(\mu, \Sigma)$. Left Column. Multivariate Gaussian $\mathcal{N}(\mu, \Sigma)$ with $\mu_X = \mu_Y = 0$, $\sigma_X = \sigma_Y = 1$, and $\sigma_{XY} = \sigma_{YX} = 0$. Right Column. Multivariate Gaussian $\mathcal{N}(\mu, \Sigma)$ with $\mu_X = \mu_Y = 0$, $\sigma_X = \sigma_Y = 1$, and $\sigma_{XY} = \sigma_{YX} = .6$. We present the raw data (top), the data-intensive framework (middle), and the framework profiles (bottom). Our data-intensive framework uses a $E$ data binning with $n$ equals to 30 and $\text{max}$ equals to $99^{th}$ percentile for both $X$ and $Y$; the $\text{min}$ values of $X$ and $Y$ are set to $0^{th}$ and $10^{th}$ percentiles, respectively.

4.1 Correlation between distance, ischemic time, and waiting time for kidneys (left) and hearts (right) using Pearson correlation. The correlation between ischemic time and distance is higher for hearts.
4.2 Raw data visualization of the association of graft failure with distance, ischemic time, and waiting time for kidneys and hearts. It is important to note that the data visualization is employing a level of transparency to allow us to distinguish between the locations where there is a high incidence of data points from the locations where the data is sparse.

4.3 Association of early graft failures and distance for kidneys (top) and hearts (bottom) using our nonparametric probabilistic data-intensive framework (Section 3.3). The distance $D$ (x-axis) ranges from 0 to 99\textsuperscript{th} distance percentile $d_{99\text{th}}$ for each organ using 10 miles bins. The graft failure time $T$ in days (y-axis) ranges from 0 (i.e., grafts that failed in the same day they were transplanted) to 30 days. (Left column) Data visualization in terms of a heat map, where each cell represents $\Pr(T \leq t|D > d)$ as estimated by Equation 3.5. (Right column) Data visualization in terms of curves, where the y-axis represents $\Pr(T \leq t|D > d)$ for 5 different values of $t \{0, 1, 7, 15, 30\}$. One can consult Table 4.1 for more information.

4.4 Association of graft failures and distance for kidneys (left) and hearts (right), considering organs that failed in at most 1 month (top) and 1 year (bottom). The distance $D$ (x-axis) exhibits 6 different values of $d \{0, 100, 200, 300, 400, 500\}$ and the y-axis represents the 95% CI for $P(T \leq t|D > d)$ as as estimated by Equation 3.5.
4.5 Association of early graft failures and ischemic time for kidneys (top) and hearts (bottom) using our data-driven framework (Section 3.3). The ischemic time $I$ in 1 hour bins (x-axis) ranges from 0 to the 99th ischemic time percentile of each organ $i_{99}$th. The graft failure time $T$ in days (y-axis) ranges from 0 (i.e., grafts that failed in the same day they were transplanted) to 30 days. (Left Column) Data visualization of the data-intensive framework in terms of a heat map, where each cell represents $\Pr(T \leq t|I > i)$ as estimated by Equation 3.5. (Right Column) Data visualization of our data-intensive framework in terms of curves, where the y-axis represents $\Pr(T \leq t|I > i)$ for 5 different values of $t \{0, 1, 7, 15, 30\}$. One can consult Table 4.2 for more information.

4.6 Association of graft failures and ischemic time for kidneys (left) and hearts (right), considering organs that failed in at most 1 month (top) and 1 year (bottom). The ischemic time $I$ (x-axis) exhibits 7 different values of $i \{0, 1, 2, 3, 4, 5, 6\}$ and the y-axis represents the 95% CI for $\Pr(T \leq t|I > i)$ as as estimated by Equation 3.5.
4.7 Association of early graft failures and waiting time for kidneys (top) and hearts (bottom) using our nonparametric probabilist approach. The waiting time $W$ in 1 hour bins (x-axis) ranges from 0 to the 99\textsuperscript{th} waiting time percentile of each organ $i_{99}$. The graft failure time $T$ in days (y-axis) ranges from 0 (i.e., grafts that failed in the same day they were transplanted) to 30 days. (Left Column) Data visualization of the nonparametric probabilist approach in terms of a heat map, where each cell represents $\Pr(T \leq t|W > w)$ as estimated by Equation 3.5. (Right Column) Data visualization of the nonparametric probabilistic approach in terms of curves, where the $y$-axis represents $\Pr(T \leq t|W > w)$ for 5 different values of $t \{0, 1, 7, 15, 30\}$. One can consult Table 4.3 for more information. . . 59

4.8 Association of graft failures and waiting time for kidneys (left) and hearts (right), considering organs that failed in at most 1 month (top) and 1 year (bottom). The waiting time $W$ (x-axis) exhibits 4 different values of $w \{0, 30, 180, 365\}$ and the $y$-axis represents the 95\% CI for $\Pr(T \leq t|W > w)$ as as estimated by Equation 3.5. . . 60
4.9 Temporal association of graft failure with distance, ischemic time, and waiting time for kidneys. The x-axis (distance, ischemic time, and waiting time) contains 30 bins equally ranging from the 10th to the 99th percentile of each factor. (Left Column) Data visualization of the nonparametric probabilist approach in terms of a heat map, where each cell represents \( \Pr(T \leq t | W > w) \) as estimated by Equation 3.5. The graft failure time \( T \) in days (y-axis) contains 30 bins equally ranging from the 10th to the 99th percentile. (Right Column) Data visualization of the nonparametric probabilistic approach in terms of curves, the y-axis represents \( \Pr(T \leq t | W > w) \), and curves are associated, from the bottommost to the topmost curve, with the 10 different percentiles of \( t \) \{10th, 19th, 29th, 39th, 49th, 59th, 69th, 79th, 89th, 99th\}. . . . . . . . . . . . . . 64
4.10 Temporal association of graft failure with distance, ischemic time, and waiting time for hearts. The x-axis (distance, ischemic time, and waiting time) contains 30 bins equally ranging from the 10\textsuperscript{th} to the 99\textsuperscript{th} percentile of each factor. (Left Column) Data visualization of the nonparametric probabilist approach in terms of a heat map, where each cell represents $\Pr(T \leq t|W > w)$ as estimated by Equation 3.5. The graft failure time $T$ in days (y-axis) contains 30 bins equally ranging from the 10\textsuperscript{th} to the 99\textsuperscript{th} percentile. (Right Column) Data visualization of the nonparametric probabilistic approach in terms of curves, the y-axis represents $\Pr(T \leq t|W > w)$, and curves are associated, from the bottommost to the topmost curve, with the 10 different percentiles of $t \{10^{\text{th}}, 19^{\text{th}}, 29^{\text{th}}, 39^{\text{th}}, 49^{\text{th}}, 59^{\text{th}}, 69^{\text{th}}, 79^{\text{th}}, 89^{\text{th}}, 99^{\text{th}}\}$.

5.1 Association between the proportion of White and other race/ethnicities. Each cell represents $\Pr(Y \leq y|X > x)$ as estimated by Equation 3.5. Using the $P$ data binning, the x-axis is the proportion of White and contains 30 bins equally ranging from the 0\textsuperscript{th} to the 90\textsuperscript{th} percentile. Using the $P$ data binning, the y-axis is the proportion of the Y race/ethnicity and contains 30 bins equally ranging from the 0\textsuperscript{th} to the 90\textsuperscript{th} percentile.
5.2 Association between the proportion of *Black* and other race/ethnicities. Each cell represents Pr(Y ≤ y|X > x) as estimated by Equation 3.5. Using the $P$ data binning, the $x$-axis is the proportion of *Black* and contains 30 bins equally ranging from the 0$^{th}$ to the 90$^{th}$ percentile. Using the $P$ data binning, the $y$-axis is the proportion of the $Y$ race/ethnicity and contains 30 bins equally ranging from the 0$^{th}$ to the 90$^{th}$ percentile.

5.3 Association between race/ethnicities. Each curve represents a race/ethnicity. The $y$-axis represents Pr(Y ≤ $y_{50}$|X > x) as estimated by Equation 3.7, and $y_{50}$ is the 50$^{th}$ proportion percentile of race/ethnicity $Y$. Using the $P$ data binning, the $x$-axis is the proportion of the $X$ race/ethnicity and contains 10 bins equally ranging from the 0$^{th}$ to the 90$^{th}$ percentile.

5.4 Association between the proportion of *White* and levels of educational attainment. Each cell represents Pr(Y ≤ y|X > x) as estimated by Equation 3.5. Using the $P$ data binning, the $x$-axis is the proportion of *White* and contains 30 bins equally ranging from the 0$^{th}$ to the 90$^{th}$ percentile. Using the $P$ data binning, the $y$-axis is the proportion of the $Y$ educational attainment and contains 30 bins equally ranging from the 0$^{th}$ to the 90$^{th}$ percentile.

xvi
5.5 Association of race/ethnicities and education. Each curve represents an educational attainment level. The $y$-axis represents $\Pr(Y \leq y_{50}|X > x)$ as estimated by Equation 3.7, and $y_{50}$ is the 50th proportion percentile of the educational level $Y$. Using the $P$ data binning, the $x$-axis is the proportion of the $X$ race/ethnicity and contains 10 bins equally ranging from the 0th to the 90th percentile.

5.6 Association between the proportion of White and levels of income. Each cell represents $\Pr(Y \leq y|X > x)$ as estimated by Equation 3.5. Using the $P$ data binning, the $x$-axis is the proportion of White and contains 30 bins equally ranging from the 0th to the 90th percentile. Using the $P$ data binning, the $y$-axis is the proportion of the $Y$ income level and contains 30 bins equally ranging from the 0th to the 90th percentile.

5.7 Association between the proportion of Elementary and levels of income. Each cell represents $\Pr(Y \leq y|X > x)$ as estimated by Equation 3.5. Using the $P$ data binning, the $x$-axis is the proportion of Elementary and contains 30 bins equally ranging from the 0th to the 90th percentile. Using the $P$ data binning, the $y$-axis is the proportion of the $Y$ income level and contains 30 bins equally ranging from the 0th to the 90th percentile.
5.8 Association of race/ethnicities and income. Each curve represents an income level. The $y$-axis represents $\Pr(Y \leq y_{50}|X > x)$ as estimated by Equation 3.7, and $y_{50}$ is the $50^{th}$ proportion percentile of the income level $Y$. Using the $P$ data binning, the $x$-axis is the proportion of the $X$ race/ethnicity and contains 10 bins equally ranging from the $0^{th}$ to the $90^{th}$ percentile.

5.9 Association of educational and income. Each curve represents an income level. The $y$-axis represents $\Pr(Y \leq y_{50}|X > x)$ as estimated by Equation 3.7, and $y_{50}$ is the $50^{th}$ proportion percentile of the educational level $Y$. Using the $P$ data binning, the $x$-axis is the proportion of the $X$ educational level and contains 10 bins equally ranging from the $0^{th}$ to the $90^{th}$ percentile.

5.10 Association between the proportion of White, Black, and Hispanic and waiting time. Each cell represents $\Pr(W \leq w|X > x)$ as estimated by Equation 3.5. Using the $P$ data binning, the $x$-axis is the proportion of the $X$ race/ethnicity and contains 30 bins equally ranging from the $0^{th}$ to the $90^{th}$ percentile. Using the $P$ data binning, the $y$-axis is waiting time $W$ and contains 30 bins equally ranging from the $0^{th}$ to the $90^{th}$ percentile.

5.11 Association of waiting time and race/ethnicity for hearts and kidneys. Each curve represents a race/ethnicity. The $y$-axis represents $\Pr(W \leq w_{50}|X > x)$ as estimated by Equation 3.7, and $w_{50}$ is the $50^{th}$ waiting time percentile of each organ. Using the $E$ data binning, the $x$-axis is the proportion of the $X$ race/ethnicity and contains 10 bins equally ranging from the 0 to .5.
5.12 Association of waiting time and race/ethnicity for hearts and kidneys over the years. Each curve represents a time period. The \( y \)-axis represents \( \Pr(W \leq w_{50}|X > x) \) as estimated by Equation 3.7, and \( w_{50} \) is the 50\(^{th} \) waiting time percentile of each organ at each period. Using the \( E \) data binning, the \( x \)-axis is the proportion of the \( X \) race/ethnicity and contains 10 bins equally ranging from the 0 to .5. .......................... 83

5.13 Association of graft failure and race/ethnicity for hearts and kidneys over the years. Each curve represents a time period. The \( y \)-axis represents \( \Pr(T \leq t_{50}|X > x) \) as estimated by Equation 3.7, and \( t_{50} \) is the 50\(^{th} \) graft failure time percentile of each organ at each period. Using the \( E \) data binning, the \( x \)-axis is the proportion of the \( X \) race/ethnicity and contains 10 bins equally ranging from the 0 to .5. 84

5.14 Association between the proportion of \( \text{High}, \text{College}, \) and \( \text{Bachelor} \) and waiting time. Each cell represents \( \Pr(W \leq w|X > x) \) as estimated by Equation 3.5. Using the \( P \) data binning, the \( x \)-axis is the proportion of the \( X \) educational level and contains 30 bins equally ranging from the 0\(^{th} \) to the 90\(^{th} \) percentile. Using the \( P \) data binning, the \( y \)-axis is waiting time \( W \) and contains 30 bins equally ranging from the 0\(^{th} \) to the 90\(^{th} \) percentile. ............ 86
5.15 Association of waiting time and educational level for hearts and kidneys. Each curve represents an educational attainment level. The $y$-axis represents $\Pr(W \leq w_{50}|X > x)$ as estimated by Equation 3.7, and $w_{50}$ is the 50th waiting time percentile of each organ. Using the $E$ data binning, the $x$-axis is the proportion of the $X$ educational level and contains 10 bins equally ranging from the 0 to .5.

5.16 Association of waiting time and educational level for hearts and kidneys over the years. Each curve represents a time period. The $y$-axis represents $\Pr(W \leq w_{50}|X > x)$ as estimated by Equation 3.7, and $w_{50}$ is the 50th waiting time percentile of each organ at each period. Using the $E$ data binning, the $x$-axis is the proportion of the $X$ race/ethnicity and contains 10 bins equally ranging from the 0 to .5.

5.17 Association of graft failure time and educational level for hearts and kidneys over the years. Each curve represents a time period. The $y$-axis represents $\Pr(T \leq t_{50}|X > x)$ as estimated by Equation 3.7, and $t_{50}$ is the 50th graft failure time percentile of each organ at each period. Using the $E$ data binning, the $x$-axis is the proportion of the $X$ race/ethnicity and contains 10 bins equally ranging from the 0 to .5.
5.18 Association between the proportion of less than 10k, from 35k to 50k, and from 150k to 200k and waiting time. Each cell represents \( \Pr(W \leq w|X > x) \) as estimated by Equation 3.5. Using the \( P \) data binning, the \( x \)-axis is the proportion of the \( X \) income level and contains 30 bins equally ranging from the 0\(^{th}\) to the 90\(^{th}\) percentile. Using the \( P \) data binning, the \( y \)-axis is waiting time \( W \) and contains 30 bins equally ranging from the 0\(^{th}\) to the 90\(^{th}\) percentile.

5.19 Association of waiting time and income level for hearts and kidneys. Each curve represents an income level. The \( y \)-axis represents \( \Pr(W \leq w_{50}|X > x) \) as estimated by Equation 3.7, and \( w_{50} \) is the 50\(^{th}\) waiting time percentile of each organ. Using the \( E \) data binning, the \( x \)-axis is the proportion of the \( X \) educational level and contains 10 bins equally ranging from the 0 to .2.

5.20 Association of waiting time and income level for hearts and kidneys over the years. Each curve represents a time period. The \( y \)-axis represents \( \Pr(W \leq w_{50}|X > x) \) as estimated by Equation 3.7, and \( w_{50} \) is the 50\(^{th}\) waiting time percentile of each organ at each period. Using the \( E \) data binning, the \( x \)-axis is the proportion of the \( X \) race/ethnicity and contains 10 bins equally ranging from the 0 to .2.
5.21 Association of graft failure time and income level for hearts and kidneys over the years. Each curve represents a time period. The \( y \)-axis represents \( \Pr(T \leq t_{50}|X > x) \) as estimated by Equation 3.7, and \( t_{50} \) is the 50\(^{th} \) graft failure time percentile of each organ at each period. Using the \( E \) data binning, the \( x \)-axis is the proportion of the \( X \) race/ethnicity and contains 10 bins equally ranging from the 0 to .2.

5.22 Proportions of patients with different race/ethnicities from counties with more than half White patients (left) and more than half Hispanic patients (right). Each curve represents 1,000 bootstrap samples. The \( y \)-axis represents the probability of each race/ethnicity.

6.1 Pearson correlation among the total number of organ-related tweets, donor registrations, and population size at the state-level. Each correlation is calculated considering the number of donor registrations and population size of each year at the state-level, except for the number of organ-related tweets, for which we only have 1-year of data (Table 3.2).

6.2 Spatial visualization of donor registrations rates (left) and organ-related tweet rates (right) aggregated at the city-level. The data visualization presents the rates grouped into five equally-spaced bins.
6.3 The relationship between organ-related tweets and donor registrations. (Left) A Poisson regression model of donor registrations predicted by organ-related tweets after controlling for population. The observed data is shown as solid blue dots, the model is shown as orange crosses, and the residuals are shown in the inset. (Right) Distribution of the typical number of donor registration rate for each quartile of tweet rates using 10,000 bootstrap samples.

6.4 Data-intensive model of organ-related tweet rates and donation registration rates using our proposed data-intensive framework. (top-left) Overall, tweet rates and donor registration rates seem to be positively related. (top-right) However, the type of relationship appears to depend on the registration rate. (bottom-left) For registration rates less than or equal to the 10\textsuperscript{th} percentile ($r_{10\text{th}} = 140.3$), a higher tweet rate seems to indicate a higher registration rate. (bottom-right) Conversely, for registration rates less than or equal to the 80\textsuperscript{th} percentile ($r_{80\text{th}} = 383.3$), a higher tweet rate seems to indicate a higher registration rate. However, due to a small sample size, the model shows a lack of statistical significance that can be visualized as the overlapping among all confidence intervals.
6.5 Awareness characterization of the six major solid organs. Each plot shows is associated with a specific organ and shows the distribution of mentions from users who mostly mention that specific organ also mention other organs. This distribution can be seen as a row on $K$ (see equations 3.3 and 3.1). Hearts, kidneys, livers, lungs, pancreas, and intestines are depicted in red, yellow, green, blue, olive, and magenta, respectively. The histogram bars are in log scale, and the bins are ranked according to the proportion of mentions.

6.6 Geographic awareness characterization of US states. Each state is characterized by the aggregated behavior of users inhabiting that state (Equations 3.2 and 3.3). The different shapes in the distribution of mentions to organs might indicate geographic variation in awareness. Each bin indicates the intensities of attention given to a particular organ. For instance, hearts and kidneys are the first and second-most-mentioned organ in most states, which may indicate the overall “ubiquity” of these transplants [61].

6.7 Characterization of geographic variation in transplantation awareness. Each state is colored according to excessive conversations about a specific organs as measured by the relative risk - $RR$ (see Eq. 6.1). Only statistically significant relative risks are considered as shown by the insets associated with the states of Louisiana, Massachusetts, and Rhode Island.
6.8 Awareness similarity of US states with regards to organ transplantation issues. Similarity matrix of states is visualized as a heat map, for which the lower values are associated with higher similarity. The hierarchical clustering of states is based on the distance between their awareness signatures (Figure 6.6). Using the dendrogram, we can analyze the clusters at any location in the hierarchy. From the leftmost state (i.e., Nebraska) to the rightmost state (i.e., Missouri) in the similarity matrix, the states are outlining zones of organ-related conversation in the following order: livers (from Delaware to North Dakota), lungs (from Massachusetts to Wisconsin), kidneys (from New York to Virginia) and hearts (from Minnesota to California). Similarly, states without a highlighted organ tend to cluster, for instance, in the zone between New Mexico and Indiana.

6.9 Groups of users who mention different organs in a similar way. To cluster the users, we used the algorithm $k$-means with the number of clusters $k$ equals to 12. This number of clusters was empirically determined and was associated with silhouette coefficient, average cluster size and inertia 0.953, 31697.42 and 2512.27, respectively. For each cluster, we show the distribution of mentions to different organs and its relative size. Possibly, these clusters might be associated with users who play different roles in the context of organ donation.
6.10 Effectiveness measures collected during our social network intervention (SNI). The measures are the (A) Impressions, Clicks, and Page Views and their normalized versions (B) Clicks per Impression and Page Views per Impression. See Tables C.1 and C.2 for details.

6.11 Correlation between effectiveness measures. The optimization plays a key role in affecting the Impressions, Clicks, and Page Views as well as their normalized versions Clicks per Impression and Page Views per Impression. Overall, these metrics become more positively associated with the optimization.

6.12 Effectiveness of our Social Network Intervention (SNI). By considering the measures taken before the optimization as a control group, SNI intervention becomes more effective as shown by our models. (A) After the optimization, clicks per impression is .0213 higher and increases $5.9 \times 10^{-7}$ at one-unit increase in the number of impressions. (B) Similarly, after the optimization, the number of page views increases $5.9^{-07}$ at one-unit increase in the number of impressions. See Tables C.1 and C.2 for details.
A.1 Benchmarking of our data-intensive approach using a negative correlation and a hidden correlation. (Left Column) A sample of 10,000 data points from a Multivariate Gaussian $\mathcal{N}(\mu, \Sigma)$, with $\mu_X = \mu_Y = 0$, $\sigma_X = \sigma_Y = 1$, and $\sigma_{XY} = \sigma_{YX} = -.6$. (Right Column) A sample of 10,000 data points from a Multivariate Gaussian $\mathcal{N}(\mu, \Sigma)$, with $\mu_X = \mu_Y = 0$, $\sigma_X = \sigma_Y = 1$, and $\sigma_{XY} = \sigma_{YX} = .0$, as well as a sample of 1,000 from a Multivariate Gaussian $\mathcal{N}(\mu, \Sigma)$, with $\mu_X = \mu_Y = 0$, $\sigma_X = \sigma_Y = 1$, and $\sigma_{XY} = \sigma_{YX} = .6$. Our data-intensive model uses a $E$ data binning with $n$ equals to 30 and $max$ equals to 99$^{th}$ percentile for both $X$ and $Y$; the $min$ values of $X$ and $Y$ are set to 0$^{th}$ and 10$^{th}$ percentiles, respectively.
A.2 Benchmarking of our data-intensive approach using a hidden sinusoidal function and the Simpson’s paradox. (Left Column) A sample of 10,000 data points from a Multivariate Gaussian \( \mathcal{N}(\mu, \Sigma) \), with \( \mu_X = \mu_Y = 0 \), \( \sigma_X = \sigma_Y = 1 \), and \( \sigma_{XY} = \sigma_{YX} = 0 \), as well as a sample of 10,000 from a sinusoidal function \( f(x) = \sin(2\pi x) \), with \( X \) uniformly distributed from \(-1\) to \(1\).

(Right Column) A sample of 10,000 data points from a Multivariate Gaussian \( \mathcal{N}(\mu, \Sigma) \), with \( \mu_X = 3 \), \( \mu_Y = 8 \), \( \sigma_X = \sigma_Y = 1 \), and \( \sigma_{XY} = \sigma_{YX} = .6 \), as well as a sample of 10,000 data points from a Multivariate Gaussian \( \mathcal{N}(\mu, \Sigma) \), with \( \mu_X = 10 \), \( \mu_Y = 3 \), \( \sigma_X = \sigma_Y = 1 \), and \( \sigma_{XY} = \sigma_{YX} = .6 \). Our data-intensive model uses a \( E \) data binning with \( n \) equals to 30 and \( \text{max} \) equals to 99\(^{th}\) percentile for both \( X \) and \( Y \); the \( \text{min} \) values of \( X \) and \( Y \) are set to 0\(^{th}\) and 10\(^{th}\) percentiles, respectively.

B.1 Raw data visualization of the association of graft failure with distance for kidneys, hearts, livers, lungs, pancreas, and intestines. It is important to note that the data visualization is employing a level of transparency to allow us distinguish between the locations where there is a high incidence of data points from the locations where the data is sparse.
B.2 Association of early graft failures and distance for kidneys, hearts, livers, lungs, pancreas, and intestines using our nonparametric probabilist approach (3.3). The distance $D$ (x-axis) ranges from 0 to $99^{th}$ distance percentile $d_{99}$ for each organ using 10 miles bins. The graft failure time $T$ in days (y-axis) ranges from 0 (i.e., grafts that failed in the same day they were transplanted) to 30 days. Data visualization in terms of a heat map, where each cell represents $Pr(T \leq t | D > d)$ as estimated by Equation 3.5.

B.3 Association of early graft failures and distance for kidneys, hearts, livers, lungs, pancreas, and intestines using our nonparametric probabilist approach (3.3). The distance $D$ (x-axis) ranges from 0 to $99^{th}$ distance percentile $d_{99}$ for each organ using 10 miles bins. The graft failure time $T$ in days (y-axis) ranges from 0 (i.e., grafts that failed in the same day they were transplanted) to 30 days. Data visualization in terms of curves, where the y-axis represents $Pr(T \leq t | D > d)$ for 5 different values of $t \{0, 1, 7, 15, 30\}$.

B.4 Raw data visualization of the association of graft failure with ischemic time for kidneys, hearts, livers, lungs, pancreas, and intestines. It is important to note that the data visualization is employing a level of transparency to allow us distinguish between the locations where there is a high incidence of data points from the locations where the data is sparse.
B.5 Association of early graft failures and ischemic time for kidneys, hearts, livers, lungs, pancreas, and intestines using our nonparametric probabilist approach (3.3). The ischemic time $I$ in 1 hour bins (x-axis) ranges from 0 to the 99th ischemic time percentile of each organ $i_{99th}$. The graft failure time $T$ in days (y-axis) ranges from 0 (i.e., grafts that failed in the same day they were transplanted) to 30 days. Data visualization of the nonparametric probabilist approach in terms of a heat map, where each cell represents $\Pr(T \leq t | I > i)$ as estimated by Equation 3.5.  

B.6 Association of early graft failures and ischemic time for kidneys, hearts, livers, lungs, pancreas, and intestines using our nonparametric probabilist approach (3.3). The ischemic time $I$ in 1 hour bins (x-axis) ranges from 0 to the 99th ischemic time percentile of each organ $i_{99th}$. The graft failure time $T$ in days (y-axis) ranges from 0 (i.e., grafts that failed in the same day they were transplanted) to 30 days. Data visualization of the nonparametric probabilistic approach in terms of curves, where the y-axis represents $\Pr(T \leq t | I > i)$ for 5 different values of $t \{0, 1, 7, 15, 30\}$.  

B.7 Raw data visualization of the association of graft failure with waiting time for kidneys, hearts, livers, lungs, pancreas, and intestines. It is important to note that the data visualization is employing a level of transparency to allow us distinguish between the locations where there is a high incidence of data points from the locations where the data is sparse.
B.8 Association of early graft failures and waiting time for kidneys, hearts, livers, lungs, pancreas, and intestines using our nonparametric probabilist approach (3.3). The waiting time $W$ in 1 hour bins (x-axis) ranges from 0 to the 99th waiting time percentile of each organ $i_{99th}$. The graft failure time $T$ in days (y-axis) ranges from 0 (i.e., grafts that failed in the same day they were transplanted) to 30 days. Data visualization of the nonparametric probabilist approach in terms of a heat map, where each cell represents $\Pr(T \leq t|W > w)$ as estimated by Equation 3.5.

B.9 Association of early graft failures and waiting time for kidneys, hearts, livers, lungs, pancreas, and intestines using our nonparametric probabilist approach (3.3). The waiting time $W$ in 1 hour bins (x-axis) ranges from 0 to the 99th waiting time percentile of each organ $i_{99th}$. The graft failure time $T$ in days (y-axis) ranges from 0 (i.e., grafts that failed in the same day they were transplanted) to 30 days. Data visualization of the nonparametric probabilistic approach in terms of curves, where the y-axis represents $\Pr(T \leq t|W > w)$ for 5 different values of $t \{0, 1, 7, 15, 30\}$.

C.1 Demographics of the audience targeted by our Social Network Intervention (SNI) provided by Facebook. The audience was targeted based on gender, age, household ownership, and household income. Each plot shows the targeted and matched population demographics. For instance, the targeted population was successfully matched in terms of household ownership.
List of Tables

3.1 Organ transplants performed in the USA from October 1st, 1987 to approximately December 31st, 2010. .................................................. 24

3.2 Descriptive statistics of collected organ-related tweets. 975,021 tweets were collected and 134,986 are from USA users. .................... 33

4.1 Association of graft failures and distance for kidneys and hearts. Each row shows the Wilson Score 95% Confidence Intervals for the Population Proportion \( \Pr(T \leq t|D > d) \) for \( t \) equals to 30 and 365 days, the sample size \( N_{D>d} \), the number of organs that failed \( N_{T \leq t} \). 54

4.2 Association of graft failures and ischemic time for kidneys and hearts. Each row shows the Wilson Score 95% Confidence Intervals for the Population Proportion \( \Pr(T \leq t|I > i) \) for \( t \) equals to 30 and 365 days, the sample size \( N_{I>i} \), the number of organs that failed \( N_{T \leq t} \). 57

4.3 Association of graft failures and waiting time for kidneys and hearts. Each row shows the Wilson Score 95% Confidence Intervals for the Population Proportion \( \Pr(T \leq t|W > w) \) for \( t \) equals to 30 and 365 days, the sample size \( N_{W>w} \), the number of organs that failed \( N_{T \leq t} \). 61
C.1 Social Network Intervention (SNI) Engagement Effectiveness before Optimization. From Aug 4th to Sept 3rd, we daily obtained three measures provided by Facebook: the number of impressions ($I$), the number of clicks ($C$), and the number of page views ($P$). Clicks per impression ($C/I$) and page views per impression ($P/I$) are normalized versions obtained by dividing the number of clicks and page views by the number of impressions. From August 4th to August 23rd, the intervention identified among the educational contents, the one with the highest capability to increase engagement among the audience.  

C.2 Social Network Intervention (SNI) Engagement Effectiveness after Optimization. From Aug 4 to Sept 3, we obtained three measures provided by Facebook daily: the number of impressions ($I$), the number of clicks ($C$), and the number of page views ($P$). Clicks per impression ($C/I$) and page views per impression ($P/I$) are normalized versions obtained by dividing the number of clicks and page views by the number of impressions. From August 24th, the intervention only delivered the optimized content, playing a key role in exposing the most appealing content to the targeted audience. 

C.3 Effectiveness of the social network intervention using Poisson regression models. We model clicks per impression ($C/I$) and page views per impression ($P/I$) based on whether the optimization is used ($O$) as well as based on the number of impressions ($I$). We report the estimation of each coefficient, its standard error, and its significance level.
Acknowledgements

I want to acknowledge the support I received from the CAPES Foundation Science Without Borders Brazilian program grant No. 0624/2014-04, from the National Science Foundation (NSF) grant No. 1560345, from the Health Resources and Services Administration (HRSA) contract No. 234-2005-370011C, from the US Army Research Office (ARO) under Agreement Number W911NF-17-1-0127, and from the OneLegacy Foundation. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the CAPES, the NSF, the HRSA, ARO, or OneLegacy.

First of all, I want to thank my advisor Dr. Ronaldo Menezes who made himself tremendously accessible, always stimulated my creativity, and made me passionate about my research. Prof. Ronaldo made me a better human being and became a truly friend of mine.

Also, I want to thank my committee members Dr. Nezamoddin Nezamoddini-Kachouie, Dr. William Allen, Dr. Eraldo Ribeiro, Dr. Carmelo Bastos-Filho, and Dr. Martin Cadeiras for diligently evaluating all my work during all this process, putting their thoughts into my research, and providing me with their best comments and ideas. Besides the role they played as committee members,
they also effectively contributed to my knowledge and work. Dr. Nezamoddin Nezamoddini-Kachouie provided me with knowledge in mathematics and statistics while I attended his course on Mathematical Statistics 1. Dr. Eraldo Ribeiro provided me with his knowledge in mathematics and theory of probability while I attended his course on Advanced Computer Vision as well knowledge on scientific writing. Dr. Martin Cadeiras provided the organ transplantation data set and advised me with his medical expertise. All their contributions were essential to the final quality of my work.

I also want to thank Dr. Diogo Pacheco, Farag Hammamd, Dr. Marcos Oliveira, Josemar Faustino, Mahdi Seyednezhad, and Ryan Hartman for all the collaboration in my research and side projects. I want to thank the Biocomplex Lab and its members for all support I received, specially Dr. Srividhya Venugopal for helping me in the early steps with the organ transplantation data, and Dr. Hugo Barbosa who advised me to come to Florida Tech. I want to thank the friends I made during this journey: Douglas Oliveira for the warm welcome and support provided since my arrival in Melbourne; Marcos Oliveira for all amazing discussions during my Ph.D; Mahdi Seyednezhad for all emotional support and for being an amazing roommate; and Josemar Faustino for all emotional support he provided while sharing with me the same challenges during the Ph.D. Specially thanks to Diogo Pacheco for generously sharing his family, for becoming the old brother I never had and making all this journey more exciting.

Lastly, I want to thank my wife, Erika Lemos, for embracing this journey with me, and for all the emotional support, kindness, care, and patience she always provided me.
Chapter 1

Introduction

In 1954, Merrill, Murray, Harrison, and Guild transplanted the first successful human organ at the Peter Bent Brigham Hospital in Boston [49]. Sixty years later, organ transplantation has shifted from an experimental procedure to a reliable, effective, and preferred life-saving therapy for patients developing end-stage organ failure. Currently, it saves thousands of lives every year worldwide [98].

In the United States, approximately 13 thousand organs were successfully transplanted only during the first semester of 2016. Yet, about 8,000 people still die each year while waiting for a organ transplant, and 56,000 people are added to the waiting list every year¹. In the last decade, we witnessed an increase of 50% (from 80 to 120 thousand) in the length of the waiting list, whereas both the number of organs recovered for transplantation and the number of transplants performed remained stationary. The reduction of this supply-demand gap has been the target of extensive research which has focused primarily on improving

¹Based on OPTN data available at https://optn.transplant.hrsa.gov as of September 25, 2017
the efficiency of organ transplantation \cite{Rana2015,Pinheiro2016} by developing new medical technologies \cite{27, 26} and designing new policies \cite{56}.

Ideally, organ allocation should maximize efficiency by allocating organs in terms of optimizing their survivability. For instance, it is preferred to allocate organs to patients with whom the grafts have higher chances of longer survival periods (e.g., more than five years) and avoid allocating organs to patients associated with a high likelihood of an early graft failure. However, solely focusing on the improvement of efficiency can lead to undesirable disparities in organ transplantation because racial/ethnic, socio, economic, and geographic factors affects health \cite{5, 65, 66}. To prevent such disparities, considerable research has also focused on improving the equity of organ transplantation by designing new policies to ensure a more equitable distribution of organs among the population \cite{48, 42}.

However, improvements on efficiency and equity are constrained by the current imbalance between donors and recipients \cite{15}. While this imbalance can be mitigated by enhancing the population awareness regarding transplantation and donation of human organs \cite{79, 62}, little has been done to effectively characterize the population awareness about organ transplantation issues.

Finally, the characterization of organ transplantation is key for its assessment and subsequent improvement, but the aspects efficiency, equity, and awareness are mainly regarded as independent due to the lack of an unified framework capable of integrating them. Therefore, the state-of-the-art in organ transplantation still lacks a comprehensive characterization that includes the aspect of awareness and integrates the aspects of efficiency, equity, and awareness (Figure 1.1).

In this dissertation, we address the lack of a comprehensive characterization in organ transplantation (Figure 1.2). First, we propose a sensor allowing for the
The Intertwined Aspects of Organ Transplantation

Figure 1.1: In organ transplantation, the aspects of efficiency, equity, and awareness are intertwined. Currently, organ transplantation lacks a characterization of population awareness. A comprehensive characterization of organ transplantation needs to take into account all these aspects in an unified manner.

characterization of population awareness regarding transplantation issues using social media. Then, we propose a nonparametric probabilistic data-intensive framework to characterize efficiency, equity, and awareness in an unified manner.
A Comprehensive Data-Driven Characterization of Organ Transplantation

1. Social Media Sensor
   - Collect: Organ transplantation, donor, donation
   - Augment: Geolocation data
   - Filter: Country

2. Curation, Creation, and Integration
   - Organ Transplantation
   - Population Census
   - Awareness

3. Nonparametric Probabilistic Data-Intensive Framework
   - Equation: \( P(Y < x | X = x) \)

4. Comprehensive Characterization
   - Efficiency
     - How graft failure is affected by factors such as disease duration and waiting time?
   - Equity
     - How waiting time is affected by socioeconomic, education, and income?
   - Awareness
     - How the population mentions organ transplantation issues?

Figure 1.2: Overview of proposed methodology to characterize the efficiency, equity, and awareness of organ transplantation. 1. A sensor for population awareness about organ transplantation issues using social media. 2. Curation and integration of organ transplantation and population demographics data sets. 3. A nonparametric probabilistic data-intensive framework. 4. Characterizing organ transplantation with regards to efficiency, equity, and awareness.
Thesis statement

A comprehensive characterization of organ transplantation must account for the aspects of (i) *efficiency* to optimally allocate organs based on major survivability factors, (ii) *equity* to provide the population equitable access to transplantation, and (iii) *awareness* to accurately promote population health literacy regarding the transplantation and donation of organs. Nevertheless, these three aspects are not independent among themselves. Awareness impacts both efficiency and equity; more registered organ donors and donation consents can increase the efficiency and equity in organ transplantation. Similarly, efficiency and equity can impact each other depending on the extent in which ethnic/racial, socio, economic, and geographic factors are determinants of health. Lastly, both efficiency and equity can also impact awareness by changing the population health literacy regarding the reliability and fairness of organ transplantation, respectively. As a result, these three aspects must be considered jointly to provide a proper characterization of organ transplantation. Given the current availability of data and computational power, organ transplantation can be comprehensively characterized by integrating the aspects of efficiency, equity, and awareness using a data-driven approach(Figure 1.2). The data mining of online activities in social media allows for the characterization of the population awareness about organ transplantation issues, and the curation and integration of available data sets allows for further characterization and integration of these aspects. Finally, a nonparametric probabilistic data-intensive framework has the potential to characterize organ transplantation with regards to efficiency, equity, and awareness in an unified manner. To validate this thesis, this dissertation is organized in the following chapters.
Chapter 2 - presents the necessary background to this dissertation. It introduces the fundamental concepts and related works involving organ transplantation, data science, and social networking sites.

Chapter 3 - presents our proposed methodology to characterize organ transplantation. First, it presents the curation and integration of available data sets. Then, it describes our proposed sensor for capturing the population awareness about organ transplantation issues using social media. Lastly, it describes our proposed nonparametric probabilistic data-intensive framework.

Chapter 4 - characterizes organ transplantation with regards to efficiency using our proposed data-intensive framework. It examines how efficiency, defined as the probability of graft failure, is impacted by three major factors: the distance traveled by organs from the donor hospital to the transplant center, the ischemic time elapsed from the moment organs are recovered from the donors to the moment organs are transplanted into patients, and the waiting time elapsed from the moment patients are included in the waiting list to the moment they receive organ transplants.

Chapter 5 - characterizes organ transplantation with regards to equity using our proposed data-intensive approach. It examines how equity, defined as the change in the likelihood of receiving organ transplants, is impacted by race/ethnicity, education, and income at the population level.
Chapter 6 - characterizes organ transplantation with regards to awareness using our proposed social media sensor, common approaches in data science, as well as our proposed data-intensive framework. It demonstrates that our sensor captures the population awareness about transplantation issues by predicting organ donation registrations and uncovering patterns of awareness related to different organs, geographic locations, and social media users. Lastly, it presents a preliminary social network intervention using social media aiming at increasing awareness regarding organ donation.

Chapter 7 - summarizes our contributions to the comprehensive characterization of organ transplantation, highlighting their limitations and setting a research agenda for future works.

By the end of the dissertation, we published the following works directly related to this dissertation:

1 - A Data Science Approach for Quantifying Spatio-Temporal Effects to Graft Failures in Organ Transplantation [72]. In this work, we proposed our nonparametric probabilistic data-intensive framework and characterized organ transplantation with regards to efficiency (Sections 3.3, 4.1, and 4.2).

2 - Characterizing Organ Donation Awareness from Social Media [62]. In this work, we proposed our sensor for capturing population awareness using social media and characterized patterns of population awareness about organ transplantation issues (Sections 3.2 and 6.2).
3 - Network-Driven Interventions to Improve Awareness about Organ Donation Among Minorities [55]. In this work, we propose a social network intervention to increased the population awareness (Section 6.3).

Besides the aforementioned works, by the end of the dissertation, we also published the following works not directly related to this dissertation, but related to social media, data science, network science, and swarm intelligence.

1 - Characterization of Football Supporters from Twitter Conversations [63].

2 - Unveiling the Funding Space of Cultural Policy in Brazil [71].

3 - On the Effect of Tax Incentives to the Cultural Space of Co-Sponsorship in Brazil: A Network-Centric Approach [70].

4 - A Comparison of Community Detection Techniques Across Thematic Twitter Emoji Networks [33].

5 - Assessing the suitability of network community detection to available metadata using rank stability [32].

6 - Network-based Assessment of Swarm Algorithms [73].

7 - Unveiling Swarm Intelligence with Network Science [59].

8 - Better Exploration-Exploitation Pace, Better Swarm: Examining the Social Interactions [58].

9 - Communication Diversity in Particle Swarm Optimizers [57].
Chapter 2

Background

In this chapter, we provide the necessary background to support this dissertation. In Section 2.1, we introduce the fundamental concepts in the field of organ transplantation as well as related works regarding the aspects of efficiency, equity, and awareness. Then, in Section 2.2, we present the fundamental concepts involving Data Science, its relationship with organ transplantation, and related works intersecting these fields.

2.1 Organ Transplantation

In the US, the history of organ transplantation starts when Merrill, Murray, Harrison, and Guild transplanted the first human organ transplant at the Peter Bent Brigham Hospital in Boston [49]. The patient was a 24 years old man with untreatable severely atrophied kidneys. Since the patient had a twin brother, the possibility of organ transplantation was suggested. However, major organ transplantation was merely a speculation at that time although other tissues
such as human skin had already been transplanted with success between twins. Nevertheless, on 23 December 1954, a normal left kidney was recovered from the healthy twin and transplanted into the ill twin. After the transplant, the patient’s clinical condition improved, and the transplanted kidney was still functioning in excellent condition one year later.

**Efficiency**

Since the first transplant in 1956, significant research in areas such as surgical technology and transplant immunology has been done to improve the efficiency of organ transplantation [89]. Currently, organ transplantation saves thousands of lives every year worldwide [60, 4, 46, 92, 88]. In the US, organ transplantation saved more than 2 million life-years over the past 30 years [77]. Only in the first semester of 2016, approximately 13 thousand organ transplants were successfully performed in the United States (US), consolidating organ transplantation as an appropriate lifesaving therapy. Yet, we still witness a shortage of organs for transplantation, and 8,000 people still die each year while waiting for an organ transplant only in the US\(^1\).

Human organs are mainly recovered from deceased donors and allocated to recipients who are ordered in a waiting list [60]. When an organ becomes available from a deceased donor, the allocation policies such as medical urgency, expected benefit, and geographical constraints (i.e., distance between donor and recipient) are applied to people in the waiting list to select a match. To enforce allocation processes, the US has been divided into 11 administrative UNOS (United Network

\(^1\)Based on OPTN data available at https://optn.transplant.hrsa.gov as of June 3, 2016
for Organ Sharing) regions which are further divided into 69 OPOs (Organ Procurement Organizations). The primary criteria used for allocating organs to patient are patient severity status level and the location where the organ became available. Organs are first allocated to local OPOs based on the severity levels. When the severity status levels within the local OPOs are exhausted, and thus no match is present at the local level, the organs are allocated to the other OPOs at the regional level within the same UNOS region, and lastly to the OPOs at the national level (i.e., other UNOS regions).

The organ-recipient allocation allocates organs that maximize efficiency based on several clinical and non-clinical factors that can lead to organ failure after transplantation [38, 9, 78]. Multiple factors can contribute to graft failure and lead to the acute and chronic tissue rejection, including histocompatibility match, medical urgency, expected benefit, preservation time, geographical constraints, and patient history of diseases. Chronic rejection is the long-term loss of functionality whereas acute rejection is a severe reaction as early as one week after transplantation. By looking the data of transplanted organs in the USA since 1980’s, we see that at least 13% of kidneys and 29% of intestines failed in the first year after they were transplanted. The failure rate within two years of transplantation is at least 18% for kidneys and reaches to 38% for intestines, and just 55% of transplanted intestines survived more than three years.

The importance of organ allocation has led many scientists to suggest specific allocation strategies for different organs, including livers [38, 6], kidneys [78], and hearts [50]. For instance, allocation policies regard the survivability of the organ outside the human body, namely, the cold ischemic time, as an important factor since it is associated with the quality degradation of the organ. The medical
community has published works indicating that ischemic time, as well as the geographical distances separating donors and recipients, lead to increased risks of graft failure [25]. Previous works attempted to understand the network structure underlying the allocation of organs and found that the network of some organs is more disorganized than others [96]. These works focus on the assessment of the organ allocation process with regards to efficiency to better inform policymakers. More recently, a prediction of survival personalized to clusters of users with similar characteristics has been proposed to improve the efficiency of organ transplantation [100]. However, given that ethnic/racial, social, economic, and geographic factors generally play a role as determinants of health [44], solely focusing on improving efficiency can compromise equity in organ transplantation.

**Equity**

Health equity is one of the primary goals of research in healthcare and a critical aspect when evaluating health systems [2]. The World Health Organization defines health equity as “the absence of unfair and avoidable or remediable differences in health among social groups.”\(^2\) However, the lack of health equity is ubiquitous worldwide. The simplest example of lack of equity in health is the variation in life expectancy between countries [44]. For instance, the life expectancy at birth in Sierra Leone is 34 years, whereas that of Japan is 82 years. Life expectancy also varies within countries. For instance, there is a 20-year gap in the life expectancy of the most and least advantaged populations in the USA [44]. The correlation of health with geographic, ethnic/racial, social, and economic factors can create

\(^2\)http://www.who.int/social_determinants/thecommission
and intensify transplantation disparities among the population. When comparing
global activities of organ transplantation among different countries, social factors
have a more significant impact than the actual medical need [98]. The scarcity of
organs hinders the achievement of equity in organ transplantation; the shortage of
organs has worsened in the last decade, with the waiting list increasing 50% (from
80 to 120 thousand), whereas the number of transplants performed has remained
constant.

Indeed, previous work shows that disparity begins in early steps of the
transplantation process; racial disparities are found when analyzing incidence, risk
factors, and treatment of chronic kidney disease within the USA [66]. Previous
research indicates lack of equity among transplant patients with less formal
education, from racial minorities, and from rural areas [66]. When looking at
a single transplant center, for instance, black patients tend to spend a longer
time at each step of the transplantation process and are 59% less likely to receive a
transplant than the white population [65]. Also, when analyzing several transplant
centers, regions containing lower socioeconomic levels have fewer patients that were
finally placed on the kidney transplant waiting list [64]. When analyzing kidney
transplantations at a single transplant center, Black patients tend to spend a longer
time at each step of the transplantation process and are 59% less likely to receive a
transplant than White patients [66]. Also, when analyzing several transplant
centers, regions containing lower socioeconomic levels have fewer patients that
were finally placed on the kidney transplant waiting list [64].

The characterization of equity is necessary to evaluate organ transplantation
systems with regards to the equitable allocation of organs, and to suggest
policy changes [48]. In this sense, countries tend to create legislative and
regulatory frameworks to ensure that organs are equitably allocated among the population [98]. For instance, the US government created the National Organ Transplant Act (NOTA) in 1984 to assist in the “nationwide distribution of organs equitably among transplant patients” [1]. Similarly, to narrow racial disparities, new allocation policies are created. For instance, a new kidney allocation policy was effectively created in 2014 to alter the definition of waiting time in order to consider the date of the patient’s first dialysis because of its correlation with race/ethnicity [48].

**Awareness**

Different approaches have attempted to reduce the shortage of organs for transplantation. Common approaches include improving transplantation immunology [89], expanding the eligibility criteria for recovered organs [56], and reducing ethnic/racial disparities [48]. Despite all efforts achieved by previous work, the current shortage of organs for transplant still constraints potential improvements on the efficiency and equity of organ transplantation. For instance, the scarcity of Black and Hispanic donors constrains the selection of a match that maximizes efficiency because it reduces the availability of acceptable levels of histocompatibility for patients that are Blacks and Hispanics [14]. In this sense, the organ transplantation community is currently focused on raising the number of organ donors [15, 84, 79].

Population awareness regarding organ donation has been extensively studied because of its potential to increase the number of organ donors [84, 79]. For instance, the lack of awareness about the process of organ allocation by the population leads to an unwillingness to become organ donors [10]. In contrast,
families with higher awareness are more likely to consent for the donation of organs [84]. As in the case of most health issues, the lack of awareness of organ transplantation appears to be related to inadequate health education among the population, particularly those from ethnic/racial minorities [14].

In the case of organ transplantation, mass media and community-based interventions aiming at improving awareness among the population have already helped increase the number of registered donors [8]. Mass media interventions target a mass of unrelated individuals, whereas community-based interventions target groups of interrelated individuals. For instance, community-based interventions might target individuals who live in the same neighborhood and belong to the same race/ethnicity. Due to the role played by the social context of individuals in the transmission of awareness, community-based interventions are generally preferred over mass media [11].

Community based interventions are an indirect form of network intervention. Strictly, a network intervention leverages the network structure of the underlying social context of individuals to foster the prevention of diseases as well as the promotion of information and behavior [95]. For instance, by targeting individuals living in the same place and with the same background, community-based interventions increase the likelihood that these targeted individuals are connected to each other.

Most network interventions need first to map the underlying social structure of individuals which is neither a simple nor a cost-effective procedure [40]. As a result, social networking sites, also known as social media, have been used as a proxy for real social networks and have been applied especially in social network interventions [20, 29]. Most of these interventions using social media attempt
to promote health-related behavior change [43, 68]. Using social media, we can easily design interventions that are broadly scalable at a lower cost. Some of these social media interventions are developed upon health-specific social media [18, 19]; the majority employs well known social media such as Facebook [101] and Twitter [68, 67]. Facebook and Twitter are by far the most common social media sites. They have been used in diverse contexts of physical and social phenomena ranging from disaster management [41] to interventions promoting health behavior change [43].

For instance, Facebook has been used in interventions involving sexual health [12], physical activity [17, 28] and food safety [47]. Similarly, Twitter has also been used to carry out interventions related to weight-loss [94] and smoking cessation [68]. Although Facebook and Twitter provide a rich source of information, most of it comes as unstructured text and needs to be understood and characterized in order for us to get useful information. Previous work characterized conversations of organ donation on Facebook aiming at understanding how organ donation advocacy agencies can influence social media users to share messages among their personal network of contacts [7].

In the context of organ donation, social media applications range from identifying potential kidney donors [21], helping organ donation advocacy agencies to increase online social network engagement [7], and increasing donor registration rates [15]. Social media has evolved as a new tool to deal with the organ donation issue because it is cost-effective, and it has the potential to attain a higher population outreach [15]. Yet, despite considerable research on organ donation using social media [7, 21, 23, 13], little has been done to associate activities in social media with organ donation. Also, the effect of current interventions aiming
at increasing the number of organ donors is not sustained after the intervention period [15].

2.2 Data Science

Given multiple competing Hypothesis, Occam’s razor favors the simplest one. In the medical and health science, simple statistical approaches are still being effectively employed. These approaches still include risk ratios [85], odds ratio [53], and Pearson correlations [90]. Similarly, simple statistical approaches are still being effectively employed in the characterization of organ transplantation with regards to efficiency [31] and equity [48, 87].

Probably because they are simple, methods such as difference-in-differences [48], Kaplan-Meier [31], and Cox proportional hazards models [31] are still exhaustively used in the field of organ transplantation. Broadly, these methods can be parametric, semi-parametric, or fully parametric [45]. Parametric and semi-parametric approaches, such as linear regressions and Cox models, make assumptions about the data. For instance, linear regressions assume that the residuals are normally distributed, and Cox models assume that the relative risk amount groups of individuals are constant over time. Provided that their underlying assumptions are satisfied, these methods are capable of providing precise results that are easy to interpret. In case their assumptions are violated, they can allow for misleading conclusions to be drawn. These models might not be appropriate for a given data set, and we always need to verify the underlying assumptions of these class of models.
Conversely, nonparametric models such as Kaplan-Meier [31] make no assumptions regarding the underlying data. They tend to fit virtually all data sets without additional validations [45]. However, they are sensitive to smaller sample sizes. Yet, all these simple methods are still widely used in the field of organ transplantation [45, 86].

The terms data science, data mining, machine learning, big data, and data-driven are often used interchangeably mainly because they involve improving decision making through the analysis of data [74, 39]. Machine learning is the study of algorithms (e.g., decision trees, artificial neural networks) capable of learning from the data, and thus improving their performance (e.g., classification accuracy) as new data items become available [52, 36]. Conversely, data mining is part of the Knowledge Discovery in Databases (KDD) process [24]. It mainly deals with structured data and can be seen as the application of machine learning in large datasets [24, 3]. Big Data is related to the current availability of a high volume, velocity, and variety of data and all its related data-processing technologies [74]. The traditional data-processing approaches are not capable of handling data sets that are extremely large, streaming of a variety of data at high speeds. In this sense, Big Data requires new data-processing technologies that are scalable, flexible, and parallel. The Big-Data Hypothesis states that decision-making can be improved with additional data.

Despite the ongoing debate regarding its precise definitions and boundaries, we consider in this dissertation Data Science as the study of the fundamental principles, processes, and techniques to understand phenomena and improve decision making by extracting patterns from data [74]. It is a combination of computer science (i.e., computational tools and programming), statistics [22], as
well as problem-solving and domain expertise [39]. Among the data science, a myriad of papers published in the past years highlight the availability of data and computational power, and call for more data-intensive approaches [30, 34]. These works state that data-intensive applications have two main challenges: managing and processing exponentially growing volumes of data as well as reducing the data analysis cycle to allow for timely decisions.

In the context of health, a data-intensive approach is particularly important given the mass of available electronic patient records [34]. However, previous approaches are not inherently parallel and tend to become more complex as new components are added. A data-intensive approach must simplify models by developing inherently parallel approaches capable of exploiting the currently available and ever-increasing computational power. In this sense, a data-intensive approach can mitigate the “black box” or “interpretability” problem, which is the difficulty of explaining how a particular model comes to its conclusions [97]. Novel approaches to characterize organ transplantation have been proposed from Bayesian frameworks [80] to machine learning algorithms [100] and Big Data [91, 86]. Yet, to the best of our knowledge, previous approaches are not inherently parallel and still suffer from the interpretability problem as they become more complex because new components are added.
Chapter 3

Methodology

In this chapter, we describe our methodology to comprehensively characterize organ transplantation, which involves the curation and integration of available data sets, the creation of novel data sets, and the use of data-intensive framework. In Section 3.1, we describe the available data sets we curated in this dissertation. Then, in Section 3.2, we present the data set we created to characterize organ transplantation awareness using social media. The results of the characterization of awareness were published on the 2017 IEEE 33rd International Conference on Data Engineering (ICDE) [62]. Lastly, in Section 3.3, we present the nonparametric probabilistic data-intensive framework we propose in this dissertation. The results our proposed framework were partially published on the 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) [72].
3.1 Curating and Integrating Data Sets

To perform a comprehensive characterization, we use available data from organ transplants, population demographics, and donation registrations. In this section, we describe each data set along with the respective curation we performed.

Curating Organ Transplantation Data

The first data set we curated is the organ transplantation. It is a record of all official transplants performed in a country regarding the major solid organs recovered from both deceased and living donors. This data set is generally divided into patients in the waiting list, recovered organs from donors, and transplanted organs into recipients. Commonly, it also contains demographic and geographic information associated with both donors and recipients.

In the USA, transplantation data is provided by the United Network for Organ Sharing (UNOS) for the six major solid organ transplants: hearts (13.21%), intestines (0.52%), lungs (5.54%), livers (27.07%), kidneys (51.86%), and pancreas (1.81%). The transplantation data available during this dissertation was provided by the Department of Medicine at the University of California, Los Angeles. It contains 373,870 transplants from October 1st, 1987 to approximately December 31st, 2010. Later, a data set containing a most recent period was also provided, but after we started the curation, fewer variables were available.

When curating organ transplantation, we standardized the data from each organ into the following variables

1. Home address - the US zip code associated with the home address of donors and recipients.
2. *Race/Ethnicity* - the ethnicity/race of donors and recipients encoded into the following 7 categories: *White, Black, Hispanic, Indian, Asian, Pacific, and Other.*

3. *Educational Level* - the educational level of donors and recipients encoded into the following 4 categories: *Elementary, High, College, and Bachelor.*

4. *Distance* - the distance (in miles) traveled by organs from the donor hospital to the transplant center\(^1\).

5. *Waiting time* - the time (in days) elapsed since the date the patient was included in the waiting list.

6. *Ischemic time* - the time (in hours) elapsed from the moment the organ was recovered from the donor to the moment it was transplanted into the patient.

7. *Survival time* - the time (in days) elapsed from the moment the organ was transplanted into the patient to the moment it failed.

We disregarded living donation data because it represents only a small proportion of transplants (see Table 3.1). The number of patients added to the waiting list (Figure 3.1) and the number of transplants varies depending on the type of organ transplant (Figure 3.2).

---

\(^1\)We can predict this distance using linear regression \((R^2 = .9897)\). The predictor of the linear regression is the distance traveled through the road network as estimated by the Google Maps Distance Matrix API, available at https://developers.google.com/maps/documentation/distance-matrix. Yet, we only validated this regression using the kidney data set.
Figure 3.1: Patients added to the waiting list in the USA during the period of 1998 to 2012 for the six major solid organs (extracted from [60]).

Figure 3.2: Transplants performed during the year in the USA during the period of 1998 to 2012 for the six major solid organs (extracted from [60]).

**Curating Population Demographics Data**

Commonly, organ transplantation is mainly assessed with regards to equity using ethnic/racial and educational data from patients within the organ transplantation data. However, as we discussed in Section 2.1, organ transplantation data is already biased due to existing disparities prior to transplantation.
Table 3.1: Organ transplants performed in the USA from October 1st, 1987 to approximately December 31st, 2010.

<table>
<thead>
<tr>
<th>Organ</th>
<th>Overall Transplants (N = 482,195)</th>
<th>Deceased Transplants (N = 373,870)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kidneys</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>297,634 (61.72%)</td>
<td>193,877 (51.86%)</td>
</tr>
<tr>
<td>Deceased</td>
<td>193,877 (65.14%)</td>
<td></td>
</tr>
<tr>
<td>Liver</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>105,405 (21.86%)</td>
<td>101,195 (27.07%)</td>
</tr>
<tr>
<td>Deceased</td>
<td>101,195 (96.01%)</td>
<td></td>
</tr>
<tr>
<td>Hearts</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>49,436 (10.25%)</td>
<td>49,395 (13.21%)</td>
</tr>
<tr>
<td>Deceased</td>
<td>49,395 (99.92%)</td>
<td></td>
</tr>
<tr>
<td>Pancreas</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>6,787 (1.41%)</td>
<td>6,759 (1.81%)</td>
</tr>
<tr>
<td>Deceased</td>
<td>6,759 (99.59%)</td>
<td></td>
</tr>
<tr>
<td>Lungs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>20,945 (4.34%)</td>
<td>20,694 (5.54%)</td>
</tr>
<tr>
<td>Deceased</td>
<td>20,694 (98.80%)</td>
<td></td>
</tr>
<tr>
<td>Intestines</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>1,988 (0.41%)</td>
<td>1,950 (0.52%)</td>
</tr>
<tr>
<td>Deceased</td>
<td>1,950 (98.09%)</td>
<td></td>
</tr>
</tbody>
</table>

To comprehensively assess equity, we also included in this dissertation external ethnic/racial and educational data using population demographics provided by the American Community Survey (ACS)\(^2\). While the census is only collected at

\(^2\)https://www.census.gov/programs-surveys/acs
each decade, the ACS data set is collected every year and contains ethnic/racial, social, and economic data collected from the population at multiple geographic levels. Also, it is collected during a period instead of a specific point in time. For instance, the 1-year estimates contains data collected during 12 months and the 5-year estimates contains data collected during a 60-month period.

In this dissertation, we curated ethnic/racial, educational, and income data from the American Community Survey. The ethnic/racial variables divide the population into the following 7 categories: White, Black, Hispanic, Indian, Asian, Pacific, and Other. Except for Hispanics, all other ethnicity/races are considered Non-Hispanic [65, 79]. For instance, White and Black should be interpreted as Non-Hispanic White and Non-Hispanic Black, respectively. Likewise, the educational variables divide the population into the following 4 categories: Elementary, High, College, and Bachelor. Each category should be interpreted as the highest educational attainment. While College refers to the population which attended college and not necessarily finished, Bachelor refers to the population which necessarily obtained a bachelor’s degree. Considering income in the preceding 12 months, the income variables divide the population into the following 10 categories of income (in U$ dollars): less than 10k, from 10k to 15k, from 15k to 25k, from 25k to 35k, from 35k to 50k, from 50k to 75k, from 75k to 100k, from 100k to 150k, from 150k to 200k, and higher than 200k.

For all US counties as well as US zip codes, we can calculate the proportion of the population belonging to each ethnic/racial, educational, and income categories aforementioned. By examining their correlation, one can see a structured correlation among ethnicity/race and income (Figure 3.3). For instance, the higher the proportion of White, the lower the proportion of ethnicity/races other
Figure 3.3: Correlation between racial/ethnic, education, and income. The ethnic/racial variables divide the population into the following 7 categories: White, Black, Hispanic, Indian, Asian, Pacific, and Other. The educational variables divide the population into the following 4 categories: Elementary, High, College, and Bachelor. Considering its income in the preceding 12 months, the income variables divide the population into the following 10 categories in dollars: less than 10k, from 10k to 15k, from 15k to 25k, from 25k to 35k, from 35k to 50k, from 50k to 75k, from 75k to 100k, from 100k to 150k, from 150k to 200k, and higher than 200k.

than White. Also, the higher the proportion of White and Asian, the lower the proportion income categories lower than 25k. Lastly, the higher the proportion of Black, Indian, and Hispanic, the higher the proportion of income categories lower than 25k. Similarly, there is a structured correlation between education and
income. For instance, the higher the proportion of Bachelor, the higher the income categories from 75k to 100k, from 100k to 150k, from 150k to 200k, and higher than 200k.

**Curating Donation Registration Data**

Donation registration data is a collection of all registered organ donors. They are potential donors from whom organs can be effectively recovered in case of a cardiopulmonary and neurological determination of death. Despite the differences between registration and attitudes [76], donation registration can indicate population awareness. In this sense, a region with higher awareness about the donation and transplantation of organs can lead to a higher number of donation registrations. In turn, a higher number of donation registration can lead to a higher number of organs recovered.

In the US, it is challenging to obtain donation registration data because it is maintained by the Department of Motor Vehicles (DMV) of each US state in a decentralized manner. In this dissertation, we only collected donation registration at the state-level. To collect this data, we manually compiled donation registrations of six annually published reports from 2009 to 2016 by the nonprofit organization Donate Life³. Each report contains accumulative registration data of each US state. Overall, registration rates increased from 37.1% in 2009, to 45% in 2012, and finally to 54% in 2016 (Figure 3.4). For instance, the registration rate in Texas increased from 2% in 2009, to 17% in 2012, and finally to 45% in 2016. Although some states currently exhibit registration rates as higher as 92% in the case of

³https://www.donatelife.net
Figure 3.4: Donation registration data at the state-level for the years 2009, 2012, and 2016. The registration of each state is normalized by the population of each state.

Montana, 12 US states still present registration rates lower than 50%. At higher resolutions, we only obtained data of donation registrations for the county of Los Angeles, California. This data are accumulated donation registration until 2013 at the level of zip code (Figure 3.5). It was provided by the nonprofit organization OneLegacy\(^4\), which obtained the data with the DMV from California.

\(^4\)https://www.onelegacy.org
Figure 3.5: Donation registration data at the level of zip code for the county of Los Angeles, California. This data is accumulated donation registrations until 2013.

### Integrating Data Sets

Ideally, data sets should exhibit common features such as spatial, temporal, and demographic information. Depending on the compatibility of such features across data sets, additional preprocessing might be necessary to integrate them; the more compatible are these characteristics, the more straightforward is the integration of data sets. For instance, spatial information can be provided at different scales such as global coordinates, zip-code, county, city, and state levels. Also, the period of temporal information available, as well as the demographic encoding used, might differ across data sets.

We mainly integrate our data sets based on the aggregation of compatible information. In our data sets, the most compatible information is the spatial information provided at the levels of zip code, city, county, and state. Although we attempt to perform our characterizations at the highest possible resolutions,
to circumvent data sparseness, we often aggregate data spatially (e.g., aggregating zip codes into cities) and temporally (e.g., aggregating temporally lagged data).

We annotate each transplant data point with the proportion of patients belonging to each ethnic/racial and educational category. Using the home address of each patient, we aggregate the transplant data at the county level and calculate the proportion of patients for each ethnic/racial and education category. Then, we associate the proportion of each race/ethnicity and educational attainment to the transplant data point based on the home address of each patient. In this sense, patients from the same county are associated with similar proportions. Lastly, we also associate each transplantation data point with the overall population demographics at the county level obtained from the American Community Survey. In the end, for each transplant data point, we have the proportion of patients for every race/ethnicity and education attainment as well as the proportion of census population for every race/ethnicity, educational attainment, and income level.

Lastly, we make variables across data sets compatible (e.g., ethnic/racial and educational variables) by aggregating subcategories not existent in all data sets. For instance, in the ACS, we calculated the educational category *High* by aggregating the 9th to 12th grade (no diploma) and High school graduate.

### 3.2 Creating a Sensor for Organ Transplantation Awareness using Social Media

Given the challenges involved in obtaining data from donation registration and its relevance to the understanding of awareness that can lead to the increase of
organ donors, in this dissertation, we attempt to create a sensor for organ donation awareness using social media.

We decided to collect data from Twitter\textsuperscript{5} not only because it is one of the most popular social media in the USA, but mainly because it is extensively used by researchers from different fields of science and allows data collection from virtually any of its users. The data collection of our tweets can be seen as a pipeline with three steps (Figure 3.6). First, we only collect tweets containing predefined predicates (i.e., keywords) of organ donation in their text. Next, we augment each tweet by converting the self-reported location into a structured address (i.e., county, state, and country) using a geographic tool, namely, OpenStreetMap\textsuperscript{6}; the self-reported location is indicated by users in their profile. Finally, each augmented tweet is filtered to those only belonging to USA users.

In our data collection, we limit our tweets to organ-related conversations by using the Twitter Stream API with a set of keywords $Q$ (Figure 3.6(1)). Precisely, $Q$ is the Cartesian product of a set of Context words (limited to organ donation terms) and a set of Subject words (limited to organs of interest). This approach attempts to guarantee that every collected tweet in our dataset contains at least one word from Context and at least one word from Subject. Therefore, our dataset is conceived in the context of organ donation.

After one year of data collection, we collected more than 130 thousands organ-related tweets from more than 70 thousand users in the USA (Table 3.2). The

\textsuperscript{5}We collected our tweets using the public Twitter Stream API available at https://dev.twitter.com/streaming/overview..

\textsuperscript{6}www.openstreetmap.org
Figure 3.6: The pipeline of organ-related collected tweets. (1) The text of each collected tweet needs necessarily to contain at least one work from the context set and at least one word from the subject set. (2) The self-reported location of each tweet is converted into a structured location using a geographic tool such as OpenStreetMap. (3) Tweets are filtered to those belonging to users from the US.

Frequency at which organs are mentioned in social media correlates with the number of transplants in the USA (Spearman correlation, $r = .84$, $p < .05$); except for hearts, which are the most mentioned on Twitter and the third on the number
Table 3.2: Descriptive statistics of collected organ-related tweets. 975,021 tweets were collected and 134,986 are from USA users.

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start Data Collection</td>
<td>Apr 22\textsuperscript{nd} 2015</td>
</tr>
<tr>
<td>Finish Data Collection</td>
<td>May 11\textsuperscript{th} 2016</td>
</tr>
<tr>
<td>Number of Days</td>
<td>385</td>
</tr>
<tr>
<td>Tweets collected</td>
<td>134,986</td>
</tr>
<tr>
<td>Number of Users</td>
<td>71,947</td>
</tr>
<tr>
<td>Avg. Tweets / Day</td>
<td>350</td>
</tr>
<tr>
<td>Avg. Tweets / User</td>
<td>1.88</td>
</tr>
<tr>
<td>Organs mentioned / Tweet</td>
<td>1.03</td>
</tr>
<tr>
<td>Organs mentioned / User</td>
<td>1.13</td>
</tr>
</tbody>
</table>

of transplants. Although more than a thousand users mention multiple organs in their tweets, most users tend to only mention a single organ (Figure 3.7(b)).

In Twitter, the geographical location of a user is commonly identified by using GPS coordinates provided for some tweets. Although GPS coordinates are dynamic and can be more precise, only 1.4% of tweets contain GPS coordinates [54]. In our case, only 0.5% of tweets contain GPS coordinates. In this dissertation, instead of using GPS coordinates, we locate users by using a geographic tool, OpenStreetMap, to convert the self-reported location provided in their profiles to a structured address, which includes information regarding the country, state, and county of the user (Figure 3.6(3)). This method has shown to be reliable even at the county level [51].

Besides characterizing the overall activity of organ-related tweets, we also employ the characterization of entities in social media proposed by Pacheco \textit{et al} [63]. By measuring the frequency at which users mention different organs in their tweets, we can formally represent $m$ users and their respective attention to
Figure 3.7: Organ-related tweets collected by our proposed sensor. (a) Distribution of the number of users over difference organs. (b) Distribution of the number of users and tweets over the number of organs mentioned.

$n$ organs using a normalized contingency matrix $\hat{U} = [\hat{u}_{ij}]_{m \times n}$. In this matrix representation, each row fully represents a user, i.e., $\sum_{j=1}^{n} \hat{u}_{ij} = 1$. 

---

(a) Number of Twitter users per organ

(b) Number of organs mentioned per user (blue) and per tweet (red)
Following the methodology proposed by Pacheco et al [63], we also characterize the average behavior of users by aggregating them based on their most mentioned organ and the geographic region they inhabit. By aggregating users based on their most mentioned organ, we can characterize dependencies among organs such as the co-occurrence of transplantation. For instance, we can uncover whether kidneys are the secondly mentioned organs among users for which hearts are the most mentioned organs. Similarly, by aggregating users based on the geographic regions they inhabit, we can characterize differences among regions that might be possibly related to different health issues, local policies, and even levels of engagement in the donation cause. For instance, we can uncover whether kidney is mentioned in Florida more than what would be expected based on how kidneys are mentioned in other states.

As proposed by Pacheco et al, each characterization can be implemented by defining a membership-indicator matrix $L$. In the characterization of organs, the membership-indicator matrix $L = [l_{ij}]_{m \times n}$ is defined as

\[
l_{ij} = \begin{cases} 
1, & \text{if } j = \arg\max_{j} \hat{U}(i, j), \\
0, & \text{otherwise.}
\end{cases}
\]  

(3.1)

Similarly, in the characterization of regions, the membership-indicator matrix $L = [l_{ij}]_{m \times r}$ is defined as
Given the users matrix $\hat{U}$, we can derive a final aggregation matrix $K$ for each membership-indicator matrix $L$. In the aggregation based on the most mentioned organ (Equation 3.1), $K = [k_{ij}]_{n \times n}$, and its rows contain the characterization of $n$ organs. Likewise, in the aggregation based on regions (Equation 3.2), $K = [k_{ij}]_{r \times n}$, and its rows represent the characterization of $r$ regions. Formally, $K$ is given by

$$K = (L^T L)^{-1} L^T \hat{U}. \quad (3.3)$$

### 3.3 A Nonparametric Probabilistic Data-Intensive Framework

In this section, we describe our proposed data-intensive nonparametric probabilistic framework to unveiling patterns in data sets. In the characterization of efficiency, equity, and awareness, we also employ several well-known statistical tools, including correlation, regression, and bootstrapping. The framework is nonparametric because it allows the model to grow with the data instead of assuming that a finite set of parameters can describe the data. Also, the framework is probabilistic because it models the data using the probabilistic framework, including probabilistic distributions, scores, and confidence intervals (CI). Lastly, the framework is data-intensive because it benefits from the currently available and
ever-increasing computational power that allows the exhaustive computation of estimates from numerous combinations of the data. It is important to note that our data-intensive framework is not affected by the “black box” or “interpretability” problem [97] because its conclusions are simply probabilities drawn based on statistically mass of associated data.

Data-Intensive Framework

Given two random variables $X$ and $Y$ in the data, our proposed framework (Figure 3.8) involves basically three data-intensive components. First, the partitioning of $X$ and $Y$, respectively, into the set $\mathcal{X}$ and $\mathcal{Y}$ of bins according to a predefined type of data binning. Next, the estimation of the relationship $\tilde{p}(X,Y)$ between $X$ and $Y$ by sampling from $\mathcal{X}$ and $\mathcal{Y}$. Finally, the data visualization of $\tilde{p}(X,Y)$.

Data-Intensive Binning

Data binning is a statistical data pre-processing technique to group data into partitions of related values. It is the first step of our framework (Figure 3.8(1)). We can use several types of data binning, including equally-spaced, logarithmically-spaced, and percentile-spaced. All types of data binning will need a minimum ($\text{min}$) and a maximum ($\text{max}$) value defined for each variable. The minimum value ($\text{min}$) can be an arbitrary value (e.g., 0), a specific percentile (e.g., $5^{\text{th}}$ percentile), or the absolute minimum value (i.e., $0^{\text{th}}$ percentile) in the data. Similarly, the maximum value ($\text{max}$) can be an arbitrary value (e.g., 30), a specific percentile (e.g., $95^{\text{th}}$ percentile), or the maximum value (i.e., $100^{\text{th}}$ percentile) in the data.

Although other binning strategies can be used (e.g., logarithmically binning), in this dissertation we mainly use the following three types of data binning: step-
Nonparametric Probabilistic Data-Intensive Framework

1. Binning
   \[ \mathcal{X} \equiv \{x_1, x_2, \ldots, x_J\} \]
   \[ \mathcal{Y} \equiv \{y_1, y_2, \ldots, y_K\} \]

Partitioning

2. Combinations
   \[ x_1, y_k \]
   \[ x_2, y_k \]
   \[ x_3, y_k \]
   \[ x_4, y_k \]
   \[ \vdots \]
   \[ x_J, y_k \]

3. Data-Intensive Model
   \[ \Pr(Y \leq y | X > x) \]

Estimation
   \[ \tilde{P}(\mathcal{X}, \mathcal{Y}) \]

Figure 3.8: Proposed Nonparametric Probabilistic Data-Intensive Framework. The three components are (1) Data binning and Partitioning, (2) Combination and Estimation, and (3) Data Visualization. (1) The supports of \( X \) and \( Y \) are partitioned into \( \mathcal{X} \) and \( \mathcal{Y} \), respectively, using a specific types of data binning. (2) For each selected combination \((x, y)\) in \((\mathcal{X}, \mathcal{Y})\), the predefined nonparametric probabilistic model \( \Pr(Y \leq y | X > x) \) is estimated by \( \tilde{p}_{x,y} \) according to the Wilson Score as calculated in Equation 3.5 (see Figure 3.9). (3) After the numerical values of all probabilities are obtained, our data-intensive framework can be visualized using a data-visualization tool such as a heat map, where each color is associated with a specific probability value.

spaced (S), equally-spaced (E), and percentile-spaced (P). In the E data binning, bins are \( n \) values equally-spaced from \( \text{min} \) to \( \text{max} \). For \( \text{min}, \text{max}, \) and \( n \) defined as 0, 5, and 3, respectively, the E data binning is \( \{0, 2.5, 5\} \). In the S data binning, bins are values ranging from \( \text{min} \) to \( \text{max} \) at each \( s \), which is a parameter specifying the difference between two consecutive values. For \( \text{min}, \text{max}, \) and \( s \) defined as 0, 5, and 1, the S data binning is also \( \{0, 1, 2, 3, 4, 5\} \). The S data binning is a special
case of the $E$ data binning with $n$ equals to 6. Lastly, in the $P$ data binning, bins are $n$ values associated with linearly-spaced percentiles, which are determined based on the mass of data existing between $min$ and $max$. It is important to note that different variables can be sampled using different types of data binning with different minimum and maximum values.

**Data-Intensive Combination**

Once we have specified the data binning, the second step of our framework is the data-intensive estimation of combinations (Figure 3.8(2)). For each combination, we can independently estimate the relationship between $X$ and $Y$. In this dissertation, we model this relationship using the model

$$
\Pr(Y \leq y \mid X > x) = \frac{\Pr(Y \leq y, X > x)}{\Pr(X > x)},
$$

(3.4)

where $\Pr(X > x)$ is the Complementary Cumulative Distribution Function (CCDF) of $X$, which is the probability of $X$ being greater than $x$. Similarly, $\Pr(Y \leq y, X > x)$ is a joint Cumulative Distribution Function (CDF) and CCDF of $Y$ and $X$, respectively, which is the probability of $X$ being greater than $x$ and $Y$ being less than or equal to $y$.

This model jointly employs these two probabilistic tools, namely the CDF and CCDF, as a way to unveil the association of $X$ and $Y$. It is important to note that although $\Pr(X > x)$ monotonically decreases as $x$ increases and $\Pr(Y \leq x)$ monotonically decreases as $y$ increases, $\Pr(Y \leq y \mid X > x)$ not necessarily monotonically increase or decrease as $x$ increases for any $y$. Actually, it will increase or decrease depending on the underlying association existing between $X$ and $Y$. 

39
Pr(Y ≤ y|X > x) = ˜p(x,y)

\[ \tilde{p} = \hat{p} + \frac{z_{\alpha/2}^2}{2n} \]

\[ \sigma_{\tilde{p}} = \frac{\sqrt{\hat{p}(1-\hat{p})/n + z_{\alpha/2}^2/4n^2}}{1 + z_{\alpha/2}^2/n} \]

Figure 3.9: Estimation employed in our proposed data-intensive framework. (1) Definition of a nonparametric probabilistic model for ˜p(x,y). (2) Calculation of the confidence interval of the estimator. (3) Calculation of the Wilson Score as the estimator. (4) Calculation of the standard error of the estimator.

Given a combination (x, y) and a significance level \( \alpha \), we propose to estimate Pr(Y ≤ y|X > x) using the Wilson score \[99\] (Figure 3.9)

\[ \tilde{p} = \frac{\hat{p} + z_{\alpha/2}^2/2n}{1 + z_{\alpha/2}^2/n} , \]
where \( n \) is the sample size, \( \hat{p} \) is simply the proportion of data points \((X, Y)\) for which \( X > x \) and \( Y \leq y \), and \( z_{\alpha/2} \) is the \( z \)-score for the significance level \( \alpha = 0.05 \).

The standard error \( \sigma_{\hat{p}} \) of this estimator \( \hat{p} \) can be calculated as

\[
\sigma_{\hat{p}} = \sqrt{\hat{p}\hat{q}/n + z^2_{\alpha/2}/4n^2},
\]

where \( \hat{q} = 1 - \hat{p} \), and the 95% confidence interval (CI) which is calculated as

\[
\hat{p} \pm z_{\alpha/2} \cdot \sigma_{\hat{p}}.
\]

It is important to note the Wilson score takes into account the skewness in the distribution of \( \hat{p} \), especially when \( p \) is close to 0 or 1, and adjust the estimator to guarantee the desired coverage probability with a significance level of \( \alpha \). For a higher sample size, the Wilson score becomes similar to the standard estimator of the population proportion \( \hat{p} \) because \( z^2_{\alpha/2}/n \), \( z^2_{\alpha/2}/2n \), and \( z^2_{\alpha/2}/4n^2 \) become negligible as \( n \to \infty \). In the case of our chosen model, the number of data points for which \( X > x \) decreases as we increase the value of \( x \).

Figure 3.10: Wilson score estimator for three different estimates of population proportion \( \hat{p} \), dashed-black): 0.1 (left), 0.5 (center), and 0.9 (right). Each plot shows the Wilson estimator \( \hat{p} \), green), its lower confidence interval \( \hat{p} + z_{\alpha/2} \cdot \sigma_{\hat{p}} \), red), and its upper confidence interval \( \hat{p} - z_{\alpha/2} \cdot \sigma_{\hat{p}} \), yellow) for sample size values from 1 to 30.
Confidence Intervals using a Data-Intensive Framework

1. Binning
   \[ \mathcal{X} = \{x_1, x_2, ..., x_J\} \]
   \[ \mathcal{Y} = \{y_1, y_2, ..., y_K\} \]

2. Sampling
   \[ x_1, y_K \]
   \[ x_2, y_K \]
   \[ x_3, y_K \]
   \[ x_4, y_K \]
   \[ \vdots \]
   \[ x_J, y_K \]
   \[ y_1 \]
   \[ y_2 \]
   \[ y_3 \]
   \[ y_4 \]
   \[ \vdots \]
   \[ y_K \]

3. 95% Confidence Intervals

\[ \Pr(Y \leq y | X > x) \]

\[ y_2 \]

\[ \mathcal{X}' \]

\[ \hat{P}(\mathcal{X}', y_2) \]

Figure 3.11: Deep examination of the association between \( X \) and \( Y \) using our proposed data-intensive framework. The three components are (1) Data binning and Partitioning, (2) Combination and Estimation, and (3) Data Visualization. (1) the supports of \( X \) and \( Y \) are partitioned into \( \mathcal{X}' \) and \( \mathcal{Y}' \), respectively, using a specific types of data binning. (2) For each selected combination \((x, y_2)\) in \((\mathcal{X}', y_2)\), the predefined nonparametric probabilistic model \( \Pr(Y \leq y | X > x) \) is estimated by \( \hat{p}_{x,y_2} \) according to the Wilson Score as calculated in Equation 3.5 (see Figure 3.9). (3) After the numerical values of all probabilities are obtained, their confidence intervals can be visualized using a data-visualization tool such as a error bar, where each bar identifies the lower and upper values of a confidence interval.

Several forms of combinations can be obtained, including the Cartesian product of \( \mathcal{X} \) and \( \mathcal{Y} \) (Figure 3.8(3)) (Figure 3.8(3)). Nevertheless, the framework is flexible enough to allow for other forms of combinations by considering \( \mathcal{X}' \) as an arbitrary subset of \( \mathcal{X} \) and \( \mathcal{Y}' \) as an arbitrary subset of \( \mathcal{X} \). We can further explore the relationship of \( X \) and \( Y \) by examining only combinations \((x, y) \in (\mathcal{X}', y_2)\) (Figure 3.11), where \( \mathcal{X}' \) can be an arbitrary subset of \( \mathcal{X} \), and \( y_2 \) an arbitrary value of
\( \mathcal{Y} \). For instance, we can infer whether \( \Pr(Y \leq y_2 | X > x_2) \) is greater than \( \Pr(Y \leq y_2 | X > x_4) \) and calculate confidence intervals to assess the statistical significance of the inference (Equation 3.7). Lastly, we can explore the existence of different relationships between \( X \) and \( Y \) by deeply examining combinations \( (x, y) \in (\mathcal{X}, \mathcal{Y}') \). In this sense, we can have different profiles of the framework over all \( x \in \mathcal{X} \) for each \( y \in \mathcal{Y}' \)

Model Profiles using a Data-Intensive Framework

1. **Binning**
   \( \mathcal{X} \equiv \{x_1, x_2, \ldots, x_J\} \)
   \( \mathcal{Y} \equiv \{y_1, y_2, \ldots, y_K\} \)

2. **Sampling**

3. **Model Profiles**

   \( \Pr(Y \leq y | X > x) \)

\( \mathcal{X} \)

\( x_1 \ x_2 \ x_3 \ x_4 \ \cdots \ x_J \)

\( \mathcal{Y}' \)

\( x_1, y_k \ x_2, y_k \ x_3, y_k \ x_4, y_k \ \cdots \ x_J, y_k \)

\( x_1, y_1 \ x_2, y_4 \ x_3, y_4 \ x_4, y_4 \ \cdots \ x_J, y_4 \)

\( x_1, y_1 \ x_2, y_2 \ x_3, y_2 \ x_4, y_2 \ \cdots \ x_J, y_2 \)

\( x_1, y_1 \ x_2, y_1 \ x_3, y_1 \ x_4, y_1 \ \cdots \ x_J, y_1 \)

Figure 3.12: Profiles of our proposed data-intensive framework. The three components are (1) Data binning and Partitioning, (2) Combination and Estimation, and (3) Data Visualization. (1) the supports of \( X \) and \( Y \) are partitioned into \( \mathcal{X} \) and \( \mathcal{Y} \), respectively, using a specific types of data binning. (2) For each selected combination \((x, y)\) in \((\mathcal{X}, \mathcal{Y}')\), the predefined nonparametric probabilistic model \( \Pr(Y \leq y | X > x) \) is estimated by \( \tilde{p}_{x,y} \) according to the Wilson Score as calculated in Equation 3.5 (see Figure 3.9). (3) After the numerical values of all probabilities are obtained, the model profile for each \( y \) can be visualized using a data-visualization tool such as multiple curves, where each curve is associated with a specific \( y \).
Benchmarking

To gain insights and start validating our data-intensive framework, we employed well-known associations between \( X \) and \( Y \) as ground truth. One of the most common models to represent the association between random variables is the Multivariate Gaussian \( \mathcal{N}(\mu, \Sigma) \), where the two parameters \( \mu \) and \( \Sigma \) are the mean and the standard deviation, respectively.

For two variables \( X \) and \( Y \), \( \mu \) is \( \mu = \begin{bmatrix} \mu_X & \mu_Y \end{bmatrix} \), where \( \mu_X \) is the mean of \( X \) and \( \mu_Y \) is the mean of \( Y \). Similarly, \( \Sigma = \begin{bmatrix} \sigma_X & \sigma_{XY} \\ \sigma_{YX} & \sigma_Y \end{bmatrix} \), where \( \sigma_X \) is the variance of \( X \), \( \sigma_Y \) is the variance of \( Y \), and \( \sigma_{XY} = \sigma_{YX} \) is the covariance between \( X \) and \( Y \). In this benchmarking, our data-intensive framework uses an \( E \) data binning with \( n \) equals to 30 and \( \text{max} \) equals to 99\(^{th}\) percentile for both \( X \) and \( Y \); the \( \text{min} \) values of \( X \) and \( Y \) are set to 0\(^{th}\) and 10\(^{th}\) percentiles, respectively.

A Multivariate Gaussian with \( \mu_X = \mu_Y = 0 \), \( \sigma_X = \sigma_Y = 1 \), and \( \sigma_{XY} = \sigma_{YX} = 0 \) can be used to model the case where \( X \) has no relationship with \( Y \) (Figure 3.13, left). In such case, the scattering of \( X \) and \( Y \) resembles a sphere (Figure 3.13, top-left), and our data-intensive framework reflects this lack of association by showing equal probability of \( Y \) values less than or equal to \( y \) for all \( x \) and \( y \) (Figure 3.13, middle-left and bottom-left).

Similarly, a Multivariate Gaussian with \( \mu_X = \mu_Y = 0 \), \( \sigma_X = \sigma_Y = 1 \), and \( \sigma_{XY} = \sigma_{YX} = .6 \) can be used to model the case where \( X \) is linearly related to \( Y \) (Figure 3.13, right). As \( X \) ranges from \(-4 \) to \( 4 \), as expected, the scattering of \( X \) and \( Y \) resembles a straight line with a positive slope (Figure 3.13, top-right). Our data-intensive framework reflects this relationship by correctly inferring, for all values of \( y \), a decreased likelihood of \( Y \) values less than or equal to \( y \) (i.e.,
increased likelihood of $Y$ values greater than $y$) (Figure 3.13, middle-right, and bottom-right). See Appendix A for more benchmarking associations.
Figure 3.13: Benchmarking of our data-intensive framework using a 10,000 data-points sampled from a Multivariate Gaussian $\mathcal{N}(\mu, \Sigma)$. **Left Column.** Multivariate Gaussian $\mathcal{N}(\mu, \Sigma)$ with $\mu_X = \mu_Y = 0$, $\sigma_X = \sigma_Y = 1$, and $\sigma_{XY} = \sigma_{YX} = 0$. **Right Column.** Multivariate Gaussian $\mathcal{N}(\mu, \Sigma)$ with $\mu_X = \mu_Y = 0$, $\sigma_X = \sigma_Y = 1$, and $\sigma_{XY} = \sigma_{YX} = .6$. We present the raw data (top), the data-intensive framework (middle), and the framework profiles (bottom). Our data-intensive framework uses a $E$ data binning with $n$ equals to 30 and $max$ equals to $99^{th}$ percentile for both $X$ and $Y$; the $min$ values of $X$ and $Y$ are set to $0^{th}$ and $10^{th}$ percentiles, respectively.
Chapter 4

On the Efficiency of Organ Transplantation

In this chapter, we study the aspect of efficiency in organ transplantation by considering efficiency as the probability of graft failure after transplantation. In this sense, a more efficient system allocates transplants with a lower likelihood of graft failure. Using our proposed nonparametric probabilistic data-intensive framework described in Section 3.3, we perform statistical analyzes to test whether graft failure $T$ (i.e., days from transplant to failure) is associated with the following factors:

1. the distance ($D$) that traveled by donated organs from the donor hospital to the transplant center (in miles). For the factor distance, our framework is instantiated as follows

$$P_{DT} = Pr(T \leq t | D > d) = \frac{Pr(T \leq t, D > d)}{Pr(D > d)},$$  \hspace{1cm} (4.1)
where $\Pr(D > d)$ is the probability of transplants for which the organ traveled distances of at least $d$ miles, whereas $\Pr(T \leq t, D > d)$ is the probability of transplants for which the organ traveled distances greater than $d$ miles and failed in at most in $t$ days.

2. the \textit{ischemic time} ($I$) elapsed in hours since the organ was recovered from the donor and transplanted into the recipient. For the factor ischemic time, our framework is instantiated as follows

$$P_{IT} = \Pr(T \leq t | I > i) = \frac{\Pr(T \leq t, I > i)}{\Pr(I > i)}, \quad (4.2)$$

where $\Pr(I > i)$ is the probability of transplants for which the organ was ischemic at least $i$ hours, whereas $\Pr(T \leq t, I > i)$ is the probability of transplants for which the organ was ischemic at least $i$ hours and failed in at most $t$ days.

3. the \textit{waiting time} ($W$) elapsed in days since the patient was included in the waiting list and received a transplant. For the factor waiting time, our framework is instantiated as follows

$$P_{WT} = \Pr(T \leq t | W > w) = \frac{\Pr(T \leq t, W > w)}{\Pr(W > w)}, \quad (4.3)$$

where $\Pr(W > w)$ is the probability of transplants for which the patients waited at least $w$ days, whereas $\Pr(T \leq t, W > w)$ is the probability of transplants for which the patients waited at least $w$ days and their transplanted organ failed in at most in $t$ days.
Figure 4.1: Correlation between distance, ischemic time, and waiting time for kidneys (left) and hearts (right) using Pearson correlation. The correlation between ischemic time and distance is higher for hearts.

Usually, these factors are themselves correlated (Figure 4.1). For instance, although one could argue that ischemic time is directly related to the distance traveled by the organs, this is not the case for all organs. For organs with shorter viability time such as hearts, the ischemic time is highly correlated with the distance traveled ($r = .6, p < .01$); these organs are so critical that they cannot be ischemic for a long time, hence most of the ischemic time is due to transportation. Conversely, for organs with greater viability time such as kidneys, the correlation is not so high ($r = .3, p < .01$); usually because these organs can be ischemic for sometime in a health facility while waiting to be transported. In this sense, these factors contributing to graft failure need to be studied separately.

The association of these factors with graft failure is usually nontrivial mainly due to the complex distribution of their values. While this distribution exhibits excessively mass of data around specific locations, it also exhibits sparsity of data in other location (Figure 4.2). In the case of organ transplantation, the data
Figure 4.2: Raw data visualization of the association of graft failure with distance, ischemic time, and waiting time for kidneys and hearts. It is important to note that the data visualization is employing a level of transparency to allow us to distinguish between the locations where there is a high incidence of data points from the locations where the data is sparse.

tends to become highly sparse as we attempt to examine transplantation of organs associated with higher distances, higher ischemic times, and for which patients
waited a higher number of days. For example, kidney and heart transplants mainly
travel distances less than 1000 and 500 miles, respectively.

By using our proposed data-intensive framework, we can explore graft failure
regarding distance, ischemic time, and waiting time, separately, while accounting
for data sparsity due to the CCDF component of the framework. In this chapter,
we focus on kidneys and hearts without the loss of generality, and the results for
all six organs can be found in Appendix B.

4.1 Efficiency and Distance

We examine the association between graft failure and the distance traveled by the
organs according to the model described in Equation 4.1. By using our framework,
we can clearly distinguish how the distance differently affects the failure of kidneys
and hearts (Figure 4.3). One can see that as the distance increases, heart failures
tend to increase, whereas kidney failures seem not to be affected.

By looking closer (Figure 4.4), we can see that kidney failure actually decreases
from 7.60% (95% CI, 7.47, 7.73), for kidneys traveling at least 0 miles, to 6.69%
(95% CI, 6.41, 6.98), for kidneys traveling at least 500 miles. Although this
decrease is approximately 1%, it can potentially be associated with hundreds of
failures given the high number of kidneys. One can consult Table 4.1 for details.

Furthermore, the effect of distance can be amplified depending on the time to
failure (Figure 4.4). Comparing organ failure in 1 month with organ failure in 1
year, the effect of distance can be amplified from 1% to 2.5% for kidneys, and from
2.2% to 3.2% for hearts (Table 4.1). This increase is statistically significant and,
Figure 4.3: Association of early graft failures and distance for kidneys (top) and hearts (bottom) using our nonparametric probabilistic data-intensive framework (Section 3.3). The distance $D$ (x-axis) ranges from 0 to 99th distance percentile $d_{99th}$ for each organ using 10 miles bins. The graft failure time $T$ in days (y-axis) ranges from 0 (i.e., grafts that failed in the same day they were transplanted) to 30 days. (Left column) Data visualization in terms of a heat map, where each cell represents $Pr(T \leq t | D > d)$ as estimated by Equation 3.5. (Right column) Data visualization in terms of curves, where the y-axis represents $Pr(T \leq t | D > d)$ for 5 different values of $t \{0, 1, 7, 15, 30\}$. One can consult Table 4.1 for more information.

Although it seems relatively small, it can be associated with additional hundreds of organ failures after transplant given the large sample size.

Only focusing on the factor of distance, our framework is able to capture different dynamics on the failure of different organs. For instance, from 0 to 500 miles, heart failure tends to increase while kidney failure tends to decrease
Figure 4.4: Association of graft failures and distance for kidneys (left) and hearts (right), considering organs that failed in at most 1 month (top) and 1 year (bottom). The distance $D$ (x-axis) exhibits 6 different values of $d$ \{0, 100, 200, 300, 400, 500\} and the y-axis represents the 95% CI for $P(T \leq t | D > d)$ as estimated by Equation 3.5.

(Figure 4.4). Considering failure in at least 1 year, the probability of heart failure increases from 22.04% (95% CI, 21.65, 22.44) to 25.28% (95% CI, 23.87, 26.68) while the probability of kidney failure decreases from 22.71% (95% CI, 22.50, 22.91) to 20.21% (95% CI, 19.76, 20.67).

### 4.2 Efficiency and Ischemic Time

The ischemic time range is different for each tissue due to different viability. For instance, hearts profoundly degrade when their ischemic time reach the 4-5
Table 4.1: Association of graft failures and distance for kidneys and hearts. Each row shows the Wilson Score 95% Confidence Intervals for the Population Proportion $\Pr(T \leq t|D > d)$ for $t$ equals to 30 and 365 days, the sample size $N_{D>d}$, the number of organs that failed $N_{T\leq t}$.

<table>
<thead>
<tr>
<th>Organ</th>
<th>T</th>
<th>D</th>
<th>$N_{D&gt;d}$</th>
<th>$N_{T\leq t}$</th>
<th>95% CI (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>30</td>
<td>0</td>
<td>161748</td>
<td>12290</td>
<td><strong>7.60 (7.47, 7.73)</strong></td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>161748</td>
<td>36730</td>
<td><strong>22.71 (22.50, 22.91)</strong></td>
<td></td>
</tr>
<tr>
<td>Kidney</td>
<td>100</td>
<td>73510</td>
<td>5192</td>
<td>7.07 (6.88, 7.25)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>49218</td>
<td>3399</td>
<td>6.91 (6.69, 7.13)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>300</td>
<td>40409</td>
<td>2743</td>
<td>6.79 (6.55, 7.04)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>400</td>
<td>34798</td>
<td>2349</td>
<td>6.76 (6.49, 7.02)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>29932</td>
<td>2002</td>
<td><strong>6.69 (6.41, 6.98)</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>0</td>
<td>42847</td>
<td>3696</td>
<td><strong>8.63 (8.36, 8.90)</strong></td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>42847</td>
<td>9444</td>
<td><strong>22.04 (21.65, 22.44)</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>21470</td>
<td>2017</td>
<td>9.40 (9.01, 9.79)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>14081</td>
<td>1402</td>
<td>9.97 (9.47, 10.46)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>300</td>
<td>9762</td>
<td>980</td>
<td>10.05 (9.46, 10.65)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>400</td>
<td>6645</td>
<td>704</td>
<td>10.62 (9.88, 11.36)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>3679</td>
<td>400</td>
<td><strong>10.91 (9.91, 11.92)</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>0</td>
<td>42847</td>
<td>9444</td>
<td><strong>22.04 (21.65, 22.44)</strong></td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>42847</td>
<td>9444</td>
<td><strong>22.04 (21.65, 22.44)</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>21470</td>
<td>5123</td>
<td>23.87 (23.30, 24.44)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>14081</td>
<td>3603</td>
<td>25.59 (24.87, 26.31)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>300</td>
<td>9762</td>
<td>2538</td>
<td>26.01 (25.14, 26.88)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>400</td>
<td>6645</td>
<td>1745</td>
<td>26.27 (25.22, 27.33)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>3679</td>
<td>929</td>
<td><strong>25.28 (23.87, 26.68)</strong></td>
<td></td>
</tr>
</tbody>
</table>

hour limit. We investigate the association between graft failure and ischemic time according to the model described in Equation 4.2.

One can see that ischemic time consistently affects the failure probability of kidneys and hearts (Figure 4.5). For kidneys experiencing at least 30 hours, this
Figure 4.5: Association of early graft failures and ischemic time for kidneys (top) and hearts (bottom) using our data-driven framework (Section 3.3). The ischemic time $I$ in 1 hour bins (x-axis) ranges from 0 to the 99th ischemic time percentile of each organ $i_{99th}$. The graft failure time $T$ in days (y-axis) ranges from 0 (i.e., grafts that failed in the same day they were transplanted) to 30 days. (Left Column) Data visualization of the data-intensive framework in terms of a heat map, where each cell represents $\Pr(T \leq t | I > i)$ as estimated by Equation 3.5. (Right Column) Data visualization of our data-intensive framework in terms of curves, where the y-axis represents $\Pr(T \leq t | I > i)$ for 5 different values of $t \{0, 1, 7, 15, 30\}$. One can consult Table 4.2 for more information.

probability increases from 7.56% (95% CI, 7.44, 7.69) to 8.92% (95% CI, 8.56, 9.28)

More importantly, however, is that the model allows us to argue that ischemic time is not relevant until a certain threshold is reached. In the case of hearts, the data does not show any significantly increased risk if the organ is not ischemic.
for more than 2 hours. After that, we see that the probability of failure rapidly increases. The same can be seen in the case of other organs (Figures B.5 and B.6).

For kidneys, the threshold in which ischemic time becomes relevant seems to be around 20 hours; below that time, the ischemic time seems to be less relevant.

Using our approach, we can see that ischemic time monotonically affects hearts: the higher the ischemic time, the higher the probability of heart failure (Figure 4.6). For failure in at least one year, the probability of heart failure increases from 21.66% (95% CI, 21.29, 22.04), for at least 0 hours, to 35.34% (95% CI, 31.21, 39.46), for at least 6 hours. For kidneys, this relationship is non-monotonic. The
Table 4.2: Association of graft failures and ischemic time for kidneys and hearts. Each row shows the Wilson Score 95% Confidence Intervals for the Population Proportion Pr(T ≤ t|I > i) for t equals to 30 and 365 days, the sample size N_{I>i}, the number of organs that failed N_{T≤t}.

<table>
<thead>
<tr>
<th>Organ</th>
<th>T</th>
<th>I</th>
<th>N_{I&gt;i}</th>
<th>N_{T≤t}</th>
<th>95% CI (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kidney</td>
<td>30</td>
<td>0</td>
<td>178257</td>
<td>13481</td>
<td><strong>7.56 (7.44, 7.69)</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>10</td>
<td>154048</td>
<td>11527</td>
<td>7.48 (7.35, 7.62)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>20</td>
<td>80146</td>
<td>6207</td>
<td>7.75 (7.56, 7.93)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>30</td>
<td>24290</td>
<td>2165</td>
<td>8.92 (8.56, 9.28)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>40</td>
<td>5522</td>
<td>569</td>
<td>10.33 (9.53, 11.13)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50</td>
<td>1451</td>
<td>183</td>
<td><strong>12.71 (11.00, 14.42)</strong></td>
</tr>
<tr>
<td></td>
<td>365</td>
<td>0</td>
<td>178257</td>
<td>40547</td>
<td><strong>22.75 (22.55, 22.94)</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>10</td>
<td>154048</td>
<td>34401</td>
<td>22.33 (22.12, 22.54)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>20</td>
<td>80146</td>
<td>17704</td>
<td>22.09 (21.80, 22.38)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>30</td>
<td>24290</td>
<td>5809</td>
<td>23.92 (23.38, 24.46)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>40</td>
<td>5522</td>
<td>1502</td>
<td>27.22 (26.04, 28.39)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50</td>
<td>1451</td>
<td>474</td>
<td><strong>32.71 (30.30, 35.12)</strong></td>
</tr>
<tr>
<td>Heart</td>
<td>30</td>
<td>0</td>
<td>46254</td>
<td>3871</td>
<td><strong>8.37 (8.12, 8.62)</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>45627</td>
<td>3825</td>
<td>8.39 (8.13, 8.64)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>37289</td>
<td>3246</td>
<td>8.71 (8.42, 9.00)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>21719</td>
<td>2129</td>
<td>9.81 (9.41, 10.21)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>7163</td>
<td>858</td>
<td>12.00 (11.25, 12.75)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>1817</td>
<td>278</td>
<td>15.37 (13.72, 17.03)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6</td>
<td>511</td>
<td>85</td>
<td><strong>16.88 (13.66, 20.11)</strong></td>
</tr>
<tr>
<td></td>
<td>365</td>
<td>0</td>
<td>46254</td>
<td>10019</td>
<td><strong>21.66 (21.29, 22.04)</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>45627</td>
<td>9911</td>
<td>21.72 (21.35, 22.10)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>37289</td>
<td>8357</td>
<td>22.41 (21.99, 22.84)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>21719</td>
<td>5322</td>
<td>24.51 (23.94, 25.08)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>7163</td>
<td>1989</td>
<td>27.78 (26.74, 28.82)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>1817</td>
<td>585</td>
<td>32.23 (30.09, 34.38)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6</td>
<td>511</td>
<td>180</td>
<td><strong>35.34 (31.21, 39.46)</strong></td>
</tr>
</tbody>
</table>

The probability of kidney failure actually decreases from 22.75% (95% CI, 22.55, 22.94), for at least 0 hours of ischemic time, to 22.09% (95% CI, 21.80, 22.38) for at least 20 hours of ischemic time. Then the probability of kidney failure starts increasing.
from 23.92% (95% CI, 23.38, 24.46), for at least 30 hours of ischemic time, to 32.71% (95% CI, 30.30, 35.12), for at least 50 hours of ischemic time.

These differences might be related to underlying mechanisms of different organ allocations. It might be the case that the allocations of kidneys can be improved by a superior organ matching if kidneys are allowed to experience higher ischemic times and travel higher distances. Definitely, after a certain threshold, the tissue might suffer damages that will surpass the benefits of any superior organ matching.

## 4.3 Efficiency and Waiting Time

The third aspect we chose to examine is the waiting time. Given the current imbalance between organ donors and transplant recipients, patients who are referred to organ transplantation need to wait until an organ is available. Unfortunately, a considerable number of patients die while waiting for an organ transplant. Therefore, the time patients spend in the waiting list might affect the overall quality of their lives, and subsequently the quality of their lives after the organ transplantation is eventually performed. In this sense, we examine the association of graft failure and waiting time using our proposed approach.

We can examine how waiting time differently affects the graft failure of kidneys and hearts (Figure 4.7). Clearly, waiting time has a greater effect on early graft failure of kidneys, increasing the probability of failure from 7.50% (95% CI, 7.38, 7.62), for at least 0 days, to 7.97% (95% CI, 7.81, 8.14), for at least 365 days. Waiting time might not affect heart failures as early as 1 month, but it clearly affects for longer periods. If we examine the graft failures happening in at least 1 year, we can see that waiting time affects both kidneys and hearts.
Figure 4.7: Association of early graft failures and waiting time for kidneys (top) and hearts (bottom) using our nonparametric probabilist approach. The waiting time $W$ in 1 hour bins (x-axis) ranges from 0 to the 99th waiting time percentile of each organ $i_{99th}$. The graft failure time $T$ in days (y-axis) ranges from 0 (i.e., grafts that failed in the same day they were transplanted) to 30 days. (Left Column) Data visualization of the nonparametric probabilist approach in terms of a heat map, where each cell represents $Pr(T \leq t|W > w)$ as estimated by Equation 3.5. (Right Column) Data visualization of the nonparametric probabilistic approach in terms of curves, where the y-axis represents $Pr(T \leq t|W > w)$ for 5 different values of $t \{0, 1, 7, 15, 30\}$. One can consult Table 4.3 for more information.

From waiting time of at least 0 days to waiting time of at least 1 year, graft failure in at most 1 year increases from 22.41% (95% CI, 22.22, 22.60) to 24.10% (95% CI, 23.84, 24.36) for kidneys, while it decreases from 21.57% (95% CI, 21.20, 21.93) to 20.22% (95% CI, 19.25, 21.18) for hearts.
<table>
<thead>
<tr>
<th>Waiting Time (W) Days</th>
<th>0.074</th>
<th>0.076</th>
<th>0.078</th>
<th>0.080</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kidney One Month</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heart One Month</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kidney One Year</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heart One Year</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 4.8: Association of graft failures and waiting time for kidneys (left) and hearts (right), considering organs that failed in at most 1 month (top) and 1 year (bottom). The waiting time \( W \) (x-axis) exhibits 4 different values of \( w \) \{0, 30, 180, 365\} and the y-axis represents the 95% CI for \( \Pr(T \leq t|W > w) \) as as estimated by Equation 3.5.

### 4.4 Temporal Effects on Efficiency

The survivability of organs is clearly associated with multiple factors as we demonstrated in the previous sections. Besides, the association of these factors with graft failure depends on the type of organ. Although a specific (e.g., distance) factor might not affect graft failure as early as 1 month, it might affect graft failure within 1 year. Even when a factor is associated with graft failure, it does not necessarily mean that such factor is consistently associated with graft failures occurring at all time periods. For instance, it might be the case that a factor such
Table 4.3: Association of graft failures and waiting time for kidneys and hearts. Each row shows the Wilson Score 95% Confidence Intervals for the Population Proportion \( \Pr(T \leq t|W > w) \) for \( t \) equals to 30 and 365 days, the sample size \( N_{W > w} \), the number of organs that failed \( N_{T \leq t} \).

<table>
<thead>
<tr>
<th>Organ</th>
<th>T</th>
<th>W</th>
<th>( N_{W &gt; w} )</th>
<th>( N_{T \leq t} )</th>
<th>95% CI (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kidney</td>
<td>30</td>
<td>0</td>
<td>191776</td>
<td>14379</td>
<td>7.50 (7.38, 7.62)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>179116</td>
<td>13399</td>
<td>7.48 (7.36, 7.60)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>138523</td>
<td>10705</td>
<td>7.73 (7.59, 7.87)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>105727</td>
<td>8429</td>
<td>7.97 (7.81, 8.14)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>365</td>
<td>0</td>
<td>191776</td>
<td>42976</td>
<td>22.41 (22.22, 22.60)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>179116</td>
<td>40140</td>
<td>22.41 (22.22, 22.60)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>138523</td>
<td>31997</td>
<td>23.10 (22.88, 23.32)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>105727</td>
<td>25477</td>
<td>24.10 (23.84, 24.36)</td>
<td></td>
</tr>
<tr>
<td>Heart</td>
<td>30</td>
<td>0</td>
<td>48893</td>
<td>4125</td>
<td>8.44 (8.19, 8.69)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>34689</td>
<td>2836</td>
<td>8.18 (7.89, 8.47)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>13968</td>
<td>1199</td>
<td>8.60 (8.13, 9.06)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>6659</td>
<td>555</td>
<td>8.36 (7.69, 9.02)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>365</td>
<td>0</td>
<td>48893</td>
<td>10544</td>
<td>21.57 (21.20, 21.93)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>34689</td>
<td>7284</td>
<td>21.00 (20.57, 21.43)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>13968</td>
<td>2958</td>
<td>21.18 (20.51, 21.86)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>6659</td>
<td>1345</td>
<td>20.22 (19.25, 21.18)</td>
<td></td>
</tr>
</tbody>
</table>

as distance may play a role in graft failures as early as 1 month, but vanishes for long-term graft failures occurring in 5 years.

To unveil temporal effects on the factors affecting graft failure, we cover all support of the distribution using a data binning strategy based on percentiles instead of only focusing on failures occurring in at most 30 days. In this sense, we focus on failures occurring in equally-spaced percentiles. For instance, if we choose 10 bins, covering from the 10\(^{th}\) percentile to the 99\(^{th}\) percentile, the bins would be the values associated with the percentiles 10\(^{th}\), 19\(^{th}\), 29\(^{th}\), 39\(^{th}\), 49\(^{th}\), 59\(^{th}\), 69\(^{th}\), 79\(^{th}\), 89\(^{th}\), and 99\(^{th}\).
By profiling the model using the percentile binning strategy, one can see that the effect of distance seems to be practically negligible on kidney failures (Figure 4.9). This lack of effect seems to be consistent across failures occurring in all percentiles. Similarly, from kidney failure occurring across all percentiles, the effect of waiting time seems to be positive with a higher extent for waiting time less than 5 years and a higher extent for waiting time higher than 5 years. Lastly, the effect of ischemic time seems to be negative from 0 to 20 hours. From 20 to 50 hours, the effect of ischemic time remains negative for long-term failures (i.e., 89th and 99th percentiles), but became positive for short-term to medium-term term failures.

Conversely, the effect of distance, ischemic time, and waiting time on heart failures depends on whether it is a short-term failure, medium-term failure, or long-term failure (Figure 4.10). For distances ranging from 0 to approximately 400 miles, short-term failure moderately increases, medium-term failures rapidly increases, and long-term failures decrease. From distances ranging from 400 to 100, short-term failure increases, medium-term failure does not significantly change, and long-term failure moderately decreases.

The effect of ischemic time significantly increases heart failures across all types of failure, from short to long-term failures. The effect of waiting time on short-term failures differs from that of long-term failures. While longer waiting time increases the likelihood of long-term heart failures, it has no effect on short-term failures.
4.5 Discussions

Our proposed approach can unveil patterns of efficiency in organ transplantation. By considering efficiency as the probability of graft failure after transplantation, we demonstrate how our approach can uncover distinct patterns of graft failure for different organs, indicating significant factors associated with graft failure as well as revealing whether such association is affected by temporal effects.

Our findings contribute to the idea that allocation policies should be custom made for each organ and that perhaps even the division of the country into Organ Procurement Organizations (OPO) could also be done tailored for each organ in order to improve the allocation process [96]. The country division into OPOs and hence the direct consideration of these areas in the allocation process is likely to be stopping a better allocation of organs such as kidneys where the distance is not so relevant.

Other external factors not considered in this work need to be assessed such as contributing causes of graft failure (i.e., infection, acute rejection, and chronic rejection), the donor cause of death as well as whether the patient was noncompliant with prescribed treatments. The question whether organ allocation policies need to be individually designed for some organs still remains open. Our approach for unveiling patterns of efficiency in organ transplantation could be further examined with regards to the other aforementioned factors. Lastly, it could possibly serve as a tool to assess new organ allocation policies as well as to support the design of new policies for different organs, including with the use of computer simulations.
Figure 4.9: Temporal association of graft failure with distance, ischemic time, and waiting time for kidneys. The x-axis (distance, ischemic time, and waiting time) contains 30 bins equally ranging from the 10th to the 99th percentile of each factor. (Left Column) Data visualization of the nonparametric probabilist approach in terms of a heat map, where each cell represents \( \Pr(T \leq t|W > w) \) as estimated by Equation 3.5. The graft failure time \( T \) in days (y-axis) contains 30 bins equally ranging from the 10th to the 99th percentile. (Right Column) Data visualization of the nonparametric probabilistic approach in terms of curves, the y-axis represents \( \Pr(T \leq t|W > w) \), and curves are associated, from the bottommost to the topmost curve, with the 10 different percentiles of \( t \) \( \{10^{th}, 19^{th}, 29^{th}, 39^{th}, 49^{th}, 59^{th}, 69^{th}, 79^{th}, 89^{th}, 99^{th}\} \).
Figure 4.10: Temporal association of graft failure with distance, ischemic time, and waiting time for hearts. The x-axis (distance, ischemic time, and waiting time) contains 30 bins equally ranging from the 10th to the 99th percentile of each factor. (Left Column) Data visualization of the nonparametric probabilist approach in terms of a heat map, where each cell represents \( \Pr(T \leq t|W > w) \) as estimated by Equation 3.5. The graft failure time \( T \) in days (y-axis) contains 30 bins equally ranging from the 10th to the 99th percentile. (Right Column) Data visualization of the nonparametric probabilistic approach in terms of curves, the y-axis represents \( \Pr(T \leq t|W > w) \), and curves are associated, from the bottommost to the topmost curve, with the 10 different percentiles of \( t \) \( \{10^{th}, 19^{th}, 29^{th}, 39^{th}, 49^{th}, 59^{th}, 69^{th}, 79^{th}, 89^{th}, 99^{th}\} \).
Chapter 5

On the Equity of Organ Transplantation

In organ transplantation, we need to take into account the aspect of equity to identify and prevent undesirable disparities [64, 66, 65]. The lack of equity generally relates to geographic, social, economic, and racial/ethnic factors [65], and appears even in the early steps before the transplantation process [66]. In this chapter, we study the aspect of equity in organ transplantation by considering equity as the difference in the probability of waiting for organ transplants.

Instead of focusing on the individual level, we study equity at the population level using demographic factors. Given the association of equity with demographic factors, we first discuss the structured correlation among population demographics in Section 5.1. This discussion is important to interpret the results in the subsequent sections. After discussing the structured correlation among population demographics, we use our proposed data-intensive framework to study equity in
organ transplantation with regards to ethnicity/race in Section 5.2, educational attainment in Section 5.3, and income levels in Section 5.4.

In this chapter, our data-intensive framework is instantiated as follows

\[
P_{XY} = \Pr(Y' \leq y' | X > x) = \frac{\Pr(Y' \leq y', X > x)}{\Pr(X > x)},
\]

(5.1)

where \( X \) is the proportion of a demographic factor (e.g., White race/ethnicity of transplant patients of census population) and \( Y' \) can be waiting time (\( W \)), graft failure (\( T \)), or other demographic factor (\( Y \)).

We are more interested in showing that our proposed data-intensive framework is capable of uncovering patterns from the data that have the potential to characterize the aspect of equity in organ transplantation. Although we can exhaustively profile our data-intensive framework at any percentile, we will mainly focus on \( Y \) being at most the median (i.e., the 50\textsuperscript{th} percentile) to simplify our analyzes in this chapter. By definition, the probability of a random variable being less than or equal to the median is \( 0.5 \); however, this probability can change depending on how \( Y \) is associated with \( X \). In this sense, the higher the probability \( \Pr(W \leq w_{50}, X > x) \) of patients waiting at most the median waiting time \( w_{50} \), the more likely are patients to receive organ transplants. Similarly, the higher the probability \( \Pr(T \leq t_{50}, X > x) \) of grafts failing in at most the median failure time, the more likely are organs of failing.

The proportion of demographic factors are measured at the population level. For instance, given that the proportion of a given demographic factor \( X \) is higher than \( x \), we focus on the probability \( \Pr(W \leq w_{50}, X > x) \) of patients waiting at most the median waiting time and on the probability \( \Pr(T \leq t_{50}, X > x) \)
of graft failures occurring in at most the median failure time. Given that $X$ is the proportion of White, the probability $\Pr(W \leq w_{50}, X > x)$ of patients waiting at most the median waiting time refers to patients from counties for which the proportion of White patients (or White general population when using the American Community Survey) is greater than $x$. the proportion of White patients and Bachelor patients instead of writing the proportion of white patients and patients with a bachelors degree, respectively. In this chapter, we analyze equity at the population level because we are studying groups of patients from counties with different demographics. Therefore, we are not implying that “Black patients are less likely to receive a kidney transplant” when we state that “patients from counties for which the proportion of Black patients (or population census) is greater than $x$ are less likely to receive a kidney transplant.” Instead, we are only implying that when we focus on the subset of counties for which the proportion of Black patients (or population census) is greater than $x$, patients from these counties are less likely to receive an organ transplant regardless of their race/ethnicity. It might be the case that in these counties individual Black patients are more likely to receive an organ transplant and individual White patients are not. Thus, we want to make it clear that even when we simply write “when the proportion is greater than .5, White patients are more likely to receive heart transplants” we are referring to patients of any race/ethnicity from counties for which the proportion of White patients is greater than .5. Nevertheless, we recognize the importance of analyzes at the individual level, but we prioritize a population-level analysis given its potential to inform interventions aiming to reduce disparities.
5.1 Equity and Population Demographics

As we previously mentioned, geographic, economic, social, and ethnic/racial factors are structurally correlated (Section 3.1, Figure 3.3). The lack of equity depends on the extent to which health is correlated with these factors. For instance, the higher the proportion of White, the lower the proportion of the ethnicity/races Black ($r = -0.60, p < .001$), Hispanic ($r = -0.59, p < .001$), Asian ($r = -0.27, p < .001$), and Indian ($r = -0.29, p < .001$). Here, we use our data-intensive framework to explore the relationship among demographic factors. To account for differences in the distribution of each factor, we use the percentile $P$ data binning (Section 3.3) for the $X$ variable. For both variables, we use $n$ equals to 30 and $\max$ equals to $90^{th}$ percentile; the $\min$ values of $X$ and $Y$ are set to $0^{th}$ and $10^{th}$ percentiles, respectively.

Categories of Race/Ethnicity

We can further explore the aforementioned structured correlation between race/ethnicities. The median (i.e., $50^{th}$ percentile) proportion of White, Black, Hispanic, Asian, and Indian are .846, .021, .038, .006, and .003, respectively. As the proportion of White increases, the likelihood of Black, Hispanic, Asian, and Indian decreases (Figure 5.1). This result is expected given that the proportions of categories belonging to the same variables are dependent and must sum up to one. This relationship can be further examined by using our data-intensive framework to estimate confidence intervals (Figure 5.3). Precisely, as the proportion of White increases to values greater than the median, the likelihood of finding counties for which the proportion of Black, Hispanic, Asian, and Indian is equal to or less than
Figure 5.1: Association between the proportion of White and other race/ethnicities. Each cell represents $\text{Pr}(Y \leq y|X > x)$ as estimated by Equation 3.5. Using the $P$ data binning, the $x$-axis is the proportion of White and contains 30 bins equally ranging from the 0th to the 90th percentile. Using the $P$ data binning, the $y$-axis is the proportion of the $Y$ race/ethnicity and contains 30 bins equally ranging from the 0th to the 90th percentile.

The median changes from .5 to .738 (95% CI, .716 to .759), .734 (95% CI, .712 to .756), .674 (95% CI, .648 to .700), and .522 (95% CI, .494 to .549), respectively. For instance, given that a county has a proportion of White greater than the median, the probability that this county has also a proportion of Hispanic lower than or equal to the median is .738. Therefore, we are more likely to find a county with a lower proportion of Hispanic if that county has a higher proportion of White.

However, as the proportion on Black increases, the likelihood of White and Indian decreases while the likelihood of Hispanic and Asian increases (Figure 5.2).
Interestingly, given that county has a high proportion of Black, that county is also more likely to have a higher proportion of Hispanic and Asian. We examine the relationship between race/ethnicity by estimating confidence intervals using our data-intensive framework (Figure 5.3). Precisely, as the proportion of Black increases to greater than the median, the likelihood of counties with the proportion of White, Hispanic, Asian, and Indian equal to or less than the median changes from .5 to .826 (95% CI, .805 to .847), .42 (95% CI, .392 to .447), .393 (95% CI, .367 to .42), and .614 (95% CI, .587 to .641), respectively. For instance, given that
Figure 5.3: Association between race/ethnicities. Each curves represents a race/ethnicity. The $y$-axis represents $\Pr(Y \leq y_{50}|X > x)$ as estimated by Equation 3.7, and $y_{50}$ is the 50th proportion percentile of race/ethnicity $Y$. Using the $P$ data binning, the $x$-axis is the proportion of the $X$ race/ethnicity and contains 10 bins equally ranging from the 0th to the 90th percentile.

A county has a proportion of Black greater than the median, the probability that this county has a proportion of Hispanic lower than or equal to the median is $0.42$. Therefore, we are more likely to find a county with a higher proportion of Hispanic if that county has a higher proportion of Black.

Levels of Educational Attainment

Similarly, we can further explore the structured correlation between race/ethnicity and education. The median proportions of Elementary, High, College, and
Figure 5.4: Association between the proportion of White and levels of educational attainment. Each cell represents \( \Pr(Y \leq y | X > x) \) as estimated by Equation 3.5. Using the \( P \) data binning, the \( x \)-axis is the proportion of White and contains 30 bins equally ranging from the 0th to the 90th percentile. Using the \( P \) data binning, the \( y \)-axis is the proportion of the \( Y \) educational attainment and contains 30 bins equally ranging from the 0th to the 90th percentile.

Bachelor are .0599, .4262, .2404, and .2540, respectively. In particular, the likelihood of Elementary, High, College, and Bachelor depends on the proportion of White (Figure 5.4). We examine the relationship between race/ethnicity and education by estimating confidence intervals using our data-intensive framework (Figure 5.5). Precisely, as the proportion of White increases to greater than the median, the likelihood of counties with the proportion of Elementary, High, College, and Bachelor equal to or less than the median changes from .5 to .661 (95% CI, .635 to .687), .419 (95% CI, .392 to .446), .567 (95% CI, .54 to .594), and .47 (95%...
Figure 5.5: Association of race/ethnicities and education. Each curve represents an educational attainment level. The $y$-axis represents $\Pr(Y \leq y_{50}|X > x)$ as estimated by Equation 3.7, and $y_{50}$ is the 50th proportion percentile of the educational level $Y$. Using the $P$ data binning, the $x$-axis is the proportion of the $X$ race/ethnicity and contains 10 bins equally ranging from the 0th to the 90th percentile.

CI, .442 to .497, respectively. For instance, given that a county has a proportion of White greater than the median, the probability that this county has a proportion of Bachelor lower than or equal to the median is .47. Therefore, we are more likely to find a county with a higher proportion of Bachelor if that county also has a higher proportion of White.
Levels of Income

Lastly, income is also structurally correlated with both race/ethnicity and education. The median proportion of less than 10k, from 15k to 25k, from 35k to 50k, from 75k to 100k, and from 150k to 200k are 0.073, 0.123, 0.147, 0.119, and 0.027, respectively. In particular, the likelihood of less than 10k, from 15k to 25k, from 35k to 50k, from 75k to 100k, and from 150k to 200k depends on the proportion of White (Figure 5.6). We examine the relationship between race/ethnicity and income by
Figure 5.7: Association between the proportion of Elementary and levels of income. Each cell represents \( \Pr(Y \leq y | X > x) \) as estimated by Equation 3.5. Using the \( P \) data binning, the \( x \)-axis is the proportion of Elementary and contains 30 bins equally ranging from the 0\(^{th}\) to the 90\(^{th}\) percentile. Using the \( P \) data binning, the \( y \)-axis is the proportion of the \( Y \) income level and contains 30 bins equally ranging from the 0\(^{th}\) to the 90\(^{th}\) percentile.

estimating confidence intervals using our data-intensive framework (Figure 5.5). Precisely, as the proportion of White increases to greater than the median, the likelihood of counties with the proportion of less than 10k, from 15k to 25k, from 35k to 50k, from 75k to 100k, and from 150k to 200k equal to or less than the median changes from .5 to .6 (95% CI, .573 to .627), .5 (95% CI, .473 to .528), .402 (95% CI, .375 to .429), and .439 (95% CI, .411 to .466), and .581 (95% CI, .553 to .608) respectively. For instance, given that a county has a proportion of White greater than the median, the probability that this county has a proportion
Figure 5.8: Association of race/ethnicities and income. Each curve represents an income level. The y-axis represents $\Pr(Y \leq y_{50}|X > x)$ as estimated by Equation 3.7, and $y_{50}$ is the 50th proportion percentile of the income level $Y$. Using the $P$ data binning, the x-axis is the proportion of the $X$ race/ethnicity and contains 10 bins equally ranging from the 0th to the 90th percentile.

of less than 10k lower than or equal to the median is .6. Therefore, we are less likely to find a county with a higher proportion of less than 10k if that county has a higher proportion of White.

Similarly, level of income also depends on the likelihood of Elementary (Figure 5.7). We examine the relationship between education and income by estimating confidence intervals using our data-intensive framework (Figure 5.9). Precisely, as the proportion of Elementary increases to greater than the median, the likelihood of counties with the proportion of less than 10k, from 15k to 25k, from 35k to 50k,
Figure 5.9: Association of educational and income. Each curve represents an income level. The $y$-axis represents $\Pr(Y \leq y_{50}|X > x)$ as estimated by Equation 3.7, and $y_{50}$ is the 50$th$ proportion percentile of the educational level $Y$. Using the $P$ data binning, the $x$-axis is the proportion of the $X$ educational level and contains 10 bins equally ranging from the 0$th$ to the 90$th$ percentile.

from 75k to 100k, and from 150k to 200k equal to or less than the median changes from .5 to .283 (95% CI, .258 to .308), .299 (95% CI, .274 to .324), .485 (95% CI, .458 to .513), and .706 (95% CI, .681 to .731), and .673 (95% CI, .647 to .699) respectively. For instance, given that a county has a proportion of Elementary greater than the median, the probability that this county has a proportion of less than 10k lower than or equal to the median is .283. Therefore, counties with a higher proportion of the income level less than 10k are more likely to be associated with a higher proportion of the educational attainment Elementary.
5.2 Equity and Race/Ethnicity

Figure 5.10: Association between the proportion of White, Black, and Hispanic and waiting time. Each cell represents \( \text{Pr}(W \leq w | X > x) \) as estimated by Equation 3.5. Using the \( P \) data binning, the \( x \)-axis is the proportion of the \( X \) race/ethnicity and contains 30 bins equally ranging from the 0\(^{th}\) to the 90\(^{th}\) percentile. Using the \( P \) data binning, the \( y \)-axis is waiting time \( W \) and contains 30 bins equally ranging from the 0\(^{th}\) to the 90\(^{th}\) percentile.
Figure 5.11: Association of waiting time and race/ethnicity for hearts and kidneys. Each curve represents a race/ethnicity. The $y$-axis represents $\Pr(W \leq w_{50}|X > x)$ as estimated by Equation 3.7, and $w_{50}$ is the 50th waiting time percentile of each organ. Using the $E$ data binning, the $x$-axis is the proportion of the $X$ race/ethnicity and contains 10 bins equally ranging from the 0 to .5.

The likelihood of patients receiving organ transplants appears to depend on the proportion of race/ethnicity within counties (Figure 5.10). In counties with an increased proportion of White, patients are less likely to receive hearts and are more likely to receive kidneys. For instance, patients from counties with more than .9 White are .457 (95% CI, .447 to .467) likely to receive a heart transplant and .578 (95% CI, .573 to .583) likely to receive a kidney transplant. Likewise, patients from counties with an increased proportion of Hispanic are more likely to receive hearts and less likely to receive kidneys. For instance, patients from counties with
more than .9 Hispanic are .627 (95% CI, .573 to .68) likely to receive hearts and are .463 (95% CI, .443 to .484) likely receive kidneys.

Using our data-intensive framework, we further examine this association by estimating the confidence intervals for the probability of waiting time for each race/ethnicity (Figure 5.11). In the case of heart transplants, patients from counties with an increased proportion of Asian and Hispanic are more likely to receive organs than patients from counties with the same increased proportion of Black and White. For patients from counties with more than half of White, Black, Hispanic, Asian, and Indian, the probability of waiting at most the median time is .486 (95% CI, .481 to .49), .497 (95% CI, .48 to .51), .585 (95% CI, .558 to .61), .656 (95% CI, .537 to .775), .606 (95% CI, .489 to .724), respectively.

In the case of kidney transplants, patients from counties with an increased proportion of White are more likely to wait at most the median time for a kidney transplant than patients from counties with the same increased proportion of Black, Hispanic, Asian, and Indian. Patients from counties with an increased proportion of Asian are less likely to wait at most the median time for a kidney transplant than patients from counties with the same increased proportion of other race/ethnicities. For patients from counties with more than half of White, Black, Hispanic, Asian, and Indian, the probability of waiting at most the median time is .528 (95% CI, .526 to .531), .416 (95% CI, .409 to .423), .489 (95% CI, .477 to .5), .328 (95% CI, .29 to .366), .394 (95% CI, .356 to .431), respectively. For Hispanic and Black, this likelihood changes for depending on the proportion. Overall, these results are consistent across the proportion of transplant patients and proportion of overall census population. Our results are consistent with the literature regarding the disparity existing on kidney transplants and its association with race/ethnicity [66].
Although our analysis of equity is at the population level, this consistency can indicate that disparity observed at the individual level is also observed at the population level for the case of kidneys.

Given possible changes that can happen over time, such as policy changes in the allocation system [75], we further explored this association across different time periods (Figure 5.12). We divided the data into four periods of five years: from 1990 to 1995, from 1995 to 2000, from 2000 to 2005, and from 2005 to 2010. Overall, the results suggest that the disparities are more significant for kidneys than hearts. Patients from counties with an increased proportion of White, Black, and Hispanic are equally likely to wait at most the median time for heart transplants. However, this likelihood changes over the years for kidney transplants.

In the period from 2005 and 2010, patients from counties with an increased proportion of White are less likely to wait (.508 95% CI, .503 to .513) at most the median time for a kidney transplant when compared to the period from 1990 to 1995 (.539 95% CI, .534 to .545). Conversely, in the same period, patients from counties associated with an increased proportion of Black are more likely to wait at most the median time for a kidney transplant than in the previous time periods (.375 95% CI, .36 to .39). Overall, the likelihood of the period from 2005 to 2010 (.470 95% CI, .455 to .486) is greater than the likelihood of the period from 2000 to 2005 (.435, 95% CI .421 to .449), which is greater than the likelihoods of the periods from 1990 to 1995 (.375 95% CI, .36 to .39), and from 1995 to 2000 (.38 95% CI .366 to .394).

In the case of counties associated with an increased proportion of Hispanic, patients are more likely to wait at most the median time for a kidney transplant in the recent period. For counties with more than .22 of these patients, the likelihood
Figure 5.12: Association of waiting time and race/ethnicity for hearts and kidneys over the years. Each curve represents a time period. The y-axis represents Pr(W ≤ w50|X > x) as estimated by Equation 3.7, and w50 is the 50th waiting time percentile of each organ at each period. Using the E data binning, the x-axis is the proportion of the X race/ethnicity and contains 10 bins equally ranging from the 0 to .5.

decreased from 1990 (.427 95% CI, .415 to 439) to 2005 (.405 95% CI, .395 to .415); only from 2005 to 2010 the likelihood increased again (.499 95% CI, .487 to .511).
Figure 5.13: Association of graft failure and race/ethnicity for hearts and kidneys over the years. Each curve represents a time period. The y-axis represents $\Pr(T \leq t_{50}|X > x)$ as estimated by Equation 3.7, and $t_{50}$ is the 50th graft failure time percentile of each organ at each period. Using the $E$ data binning, the x-axis is the proportion of the X race/ethnicity and contains 10 bins equally ranging from the 0 to .5.

Moreover, in the period from 2005 to 2010, the range of Hispanic proportion with the highest likelihood of waiting at most the median time (i.e., from .22 to .33)
is also associated with the overall lowest likelihood (Figure 5.11). These changes in the likelihood of waiting at most the median time for a kidney transplant for might be related to efforts aiming at reducing ethnic/racial disparities.

Previously, our results on efficiency suggested that increased waiting time correlates with an increased likelihood of kidney failure (Section 4.3). Interestingly, the likelihood of graft failure also seems to be associated with these aforementioned changes in the likelihood of waiting time over the years (Figure 5.13). Organs transplanted into patients from counties that are less likely to receive a transplant are also more likely to fail. In the case of hearts, the likelihood for patients from counties with an increased proportion of White does not change across the periods. Despite the overlapping among confidence intervals, organs are slightly less likely to fail for counties with an increased proportion of Black. For Hispanic, hearts transplanted in recent periods are significantly less likely to fail.

In the case of kidneys, this association seems to be greater. Kidneys transplanted into patients from counties with an increased proportion of Black and Hispanic are less likely to fail in the recent periods. For instance, kidneys transplanted into patients from counties with more than .3 Black are less likely to fail in the period from 2005 to 2010 (.504 95% CI, .495 to .514) than in the period from 2000 to 2005 (.55 95% CI, .541 to .558). Similarly, kidneys transplanted into patients from counties with more than .3 Hispanic are less likely to fail in the period from 2005 to 2010 (.489 95% CI, .474 to .505) than the period from 2000 to 2005 (.566 95% CI, .553 to .579). Interestingly, these patients are also more likely to receive kidneys in this same period (Figure 5.12).
5.3 Equity and Educational Attainment

Figure 5.14: Association between the proportion of High, College, and Bachelor and waiting time. Each cell represents $\Pr(W \leq w | X > x)$ as estimated by Equation 3.5. Using the $P$ data binning, the x-axis is the proportion of the X educational level and contains 30 bins equally ranging from the 0th to the 90th percentile. Using the $P$ data binning, the y-axis is waiting time $W$ and contains 30 bins equally ranging from the 0th to the 90th percentile.
In this section, we examine equity with regards to levels of educational attainment. Using our data-intensive framework, we can see that likelihood of waiting for kidneys and hearts appears to correlate with educational attainment (Figure 5.14). To further examine this association for hearts and kidneys, we estimate the confidence intervals for each level of educational attainment (Figure 5.15) In the case of hearts, we observe a lack of association between waiting time and educational level.
Figure 5.16: Association of waiting time and educational level for hearts and kidneys over the years. Each curve represents a time period. The $y$-axis represents $\Pr(W \leq w_{50}|X > x)$ as estimated by Equation 3.7, and $w_{50}$ is the 50th waiting time percentile of each organ at each period. Using the $E$ data binning, the $x$-axis is the proportion of the $X$ race/ethnicity and contains 10 bins equally ranging from the 0 to .5.

In the case of kidneys, the likelihood of receiving a transplant changes for specific proportions of Elementary and High. In counties for which the proportion
Figure 5.17: Association of graft failure time and educational level for hearts and kidneys over the years. Each curve represents a time period. The y-axis represents $\Pr(T \leq t_{50}|X > x)$ as estimated by Equation 3.7, and $t_{50}$ is the 50th graft failure time percentile of each organ at each period. Using the $E$ data binning, the $x$-axis is the proportion of the $X$ race/ethnicity and contains 10 bins equally ranging from the 0 to .5.

of Elementary patients is greater than zero, patients are slightly less likely (.481 95% CI, .478 to .483) to receive kidneys. Also, patients are less likely to receive
kidneys (\(0.457\) 95% CI, 0.453 to 0.461) in counties for which the proportion of High patients is greater than 0.11. Lastly, patients are more likely to receive kidneys (\(0.521\) 95% CI, 0.515 to 0.527) in counties for which the proportion of Bachelor patients is greater than 0.5.

Using the population data from the American Community Survey, we observe additional changes in the likelihood of receiving kidneys. For instance, patients are more likely to receive kidneys (\(0.56\) 95% CI, 0.552 to 0.568) in counties for which the proportion of High population is greater than 0.5. This likelihood increases with the proportion of High population. The opposite happens for Bachelor. Patients are less likely to receive kidneys (\(0.413\) 95% CI, 0.405 to 0.422) in counties for which the proportion of Bachelor population is greater than 0.5. This likelihood is in contrast with the aforementioned likelihood of patients from counties for which the proportion of Bachelor patients are greater than 0.5. Clearly, counties with a higher proportion of patients with Bachelor are not necessarily the same counties with a higher proportion of population with Bachelor.

By further examining the association between waiting time and education over the years (Figure 5.16), we also observe changes in the likelihood of receiving organs. In the case of hearts, patients from counties with an increased proportion of Bachelor are more likely to receive a transplant in the period from 1995 to 2000 (\(0.515\) 95% CI, 0.491 to 0.538) than in the recent period from 2005 to 2010 (\(0.463\) 95% CI, 0.44 to 0.487). In the case of kidney transplants, the likelihood greatly changes depending on the proportion of Bachelor patients. In counties for which the proportion of Bachelor patients is greater than 0.5, patients are less likely to receive a transplant in the recent period from 2005 to 2010 (\(0.478\) 95% CI, 0.466 to 0.491) than in the period from 2000 to 2005 (\(0.54\) 95% CI, 0.528 to 0.552).
Interestingly, the likelihood of receiving kidneys seems to correlate with the likelihood of organ failure after transplantation (Figure 5.17). In counties for which the proportion of Bachelor patients is greater than .5, kidneys transplanted into patients are more likely to fail in the recent period from 2005 to 2010 (.497 95% CI, .484 to .509) than in the period from 2000 to 2005 (.446 95% CI, .434 to .457). We observe a lack of change for other levels of educational attainment.

## 5.4 Equity and Income Levels

The transplantation data lacks information regarding the income level of patients. As a result, we can only examine the association between equity and income levels using population data from the American Community Survey. We observe some indication that waiting time is associated with income levels mainly for kidneys (Figure 5.18).

To further examine this association for hearts and kidneys, we estimate the confidence intervals for each income level (Figure 5.19). In the case of hearts, we observe a lack of change in the likelihood of receiving organs. In the case of kidneys, this likelihood changes depending on the income proportion of counties. For instance, patients are less likely to receive kidneys in counties with an increased proportion of less than 10k and from 150k to 200k. Precisely, patients are less likely to receive kidneys in counties for which the proportion of Less than 10k (.454 95% CI, .448 to .461) and from 150k to 200k (.419 95% CI, .407 to .431) is greater than .11. Likewise, patients are more likely to receive kidneys in counties with an increased proportion of other income levels. Precisely, patients are more likely to receive kidneys in counties for which the proportion of from 35k to 50k (.583 95%
Figure 5.18: Association between the proportion of less than 10k, from 35k to 50k, and from 150k to 200k and waiting time. Each cell represents \( \Pr(W \leq w | X > x) \) as estimated by Equation 3.5. Using the \( P \) data binning, the \( x \)-axis is the proportion of the \( X \) income level and contains 30 bins equally ranging from the 0\(^{th}\) to the 90\(^{th}\) percentile. Using the \( P \) data binning, the \( y \)-axis is waiting time \( W \) and contains 30 bins equally ranging from the 0\(^{th}\) to the 90\(^{th}\) percentile.

CI, .543 to .624) and from 75k to 100k (.712 95% CI, .581 to .843) is greater than .2.
Figure 5.19: Association of waiting time and income level for hearts and kidneys. Each curve represents an income level. The $y$-axis represents $Pr(W \leq w_{50}|X > x)$ as estimated by Equation 3.7, and $w_{50}$ is the $50^{th}$ waiting time percentile of each organ. Using the $E$ data binning, the $x$-axis is the proportion of the $X$ educational level and contains 10 bins equally ranging from the 0 to .2.

By further examining the association between waiting time and income over the years (Figure 5.20), we also observe changes in the likelihood of receiving organs. In the case of hearts, patients from counties with an increased proportion of $Less than 10k$ population appear to be less likely to receive a transplant in the period from 2005 to 2010; however, this change is not statistically significant. This likelihood change for kidneys with a greater extent. Despite the overlapping among some confidence intervals (Figure 5.20, top-right and bottom-right), patients from counties with an increased proportion of $Less than 10k$ and from $150k$ to $200k$ are more likely to receive kidneys in the recent period from 2005 to 2010. Precisely, patients from counties for which the proportion of $Less than 10k$ is greater than .11 are more likely to receive a kidney transplant in the period from 2005 to 2010 ($0.489 \ 95\% \ CI, \ 0.475 \ to \ 0.503$) than in the periods from 1990 to 1995 ($0.434 \ 95\% \ CI, \ 0.421 \ to \ 0.448$), from 1995 to 2000 ($0.435 \ 95\% \ CI, \ 0.423 \ to \ 0.448$), and from 2000 to 2005 ($0.457 \ 95\% \ CI, \ 0.444 \ to \ 0.469$). Conversely, patients from counties for which the proportion of $from 35k to 50k$ is greater than .15 are less likely to receive a kidney
Figure 5.20: Association of waiting time and income level for hearts and kidneys over the years. Each curve represents a time period. The y-axis represents \( \Pr(W \leq w_{50}|X > x) \) as estimated by Equation 3.7, and \( w_{50} \) is the 50\(^{th}\) waiting time percentile of each organ at each period. Using the \( E \) data binning, the \( x \)-axis is the proportion of the \( X \) race/ethnicity and contains 10 bins equally ranging from the 0 to .2.

transplant in the period from 2005 to 2010 (.518 95% CI, .506 to .529)) than in
Figure 5.21: Association of graft failure time and income level for hearts and kidneys over the years. Each curve represents a time period. The y-axis represents $\Pr(T \leq t_{50}|X > x)$ as estimated by Equation 3.7, and $t_{50}$ is the 50th graft failure time percentile of each organ at each period. Using the $E$ data binning, the $x$-axis is the proportion of the $X$ race/ethnicity and contains 10 bins equally ranging from the 0 to 0.2.

the periods from 1990 to 1995 (0.574 95% CI, 0.563 to 0.586), from 1995 to 2000 (0.588 95%CI, 0.577 to 0.60), and from 2000 to 2005 (0.569 95% CI, 0.558 to 0.579).
When we analyze the likelihood of graft failure (Figure 5.21), the impact of waiting is consistent with previous results. In this sense, patients less likely to receive organs are also more likely to experience a graft failure.

5.5 Discussions

In this chapter, we characterized equity with regards to race/ethnicity, education, income. Here, we define equity based on the likelihood of receiving organ transplants. Then, we analyze this likelihood as a function of the proportion of transplant patients and the general population from different race/ethnicity, education, and income. Our analyzes have some limitations.

First, when interpreting our results regarding equity one needs to take into account the correlation among demographic factors (Section 5.1). In this sense, an increased proportion of a specific income level (e.g., from 150k to 200k) might also correlate with an increased proportion of a specific race/ethnicity (e.g., Hispanic). Also when comparing counties for which the proportion of Hispanic is greater than .4 and counties for which the proportion of White is greater than .4., we can guarantee that we exclude from both samples counties for which the proportion of these categories is less than .4. However, some counties for which the proportion of Hispanic is greater than .4 might also exhibit a proportion of White greater than .4. Therefore, as we focus on counties associated with higher proportion of a demographic factor, counties can still overlap across our comparisons. In this sense, these counties will be present in both samples.

For instance, counties with more than half of White patients (Figure 5.22, left) are associated with patients that are 14.40% (95% CI, 14.35 to 14.47) Black,
76.15% (95% CI, 76.08 to 76.22) White, 6.03% (95% CI, 5.99 to 6.06) Hispanic, 2.42% (95% CI, 2.40 to 2.43) Asian, and 0.65% (95% CI, 0.64 to 0.67) Indian. Similarly, counties with more than half of Hispanic patients (Figure 5.22, right) are associated with patients that are 6.19% (6.02 to 6.36) Black, 19.41% (19.12 to 19.73) White, 72.77% (72.29 to 73.22) Hispanic, 0.99% (0.96 to 1.02) Asian, and 0.40% (0.35 to 0.45) Indian.

Figure 5.22: Proportions of patients with different race/ethnicities from counties with more than half White patients (left) and more than half Hispanic patients (right). Each curve represents 1,000 bootstrap samples. The y-axis represents the probability of each race/ethnicity.

Previously, we stated that in this chapter we analyze equity at the population level, and thus we are not making inferences at the individual level. Although inferences regarding counties with more than .9 White patients almost surely correlate with individual White patients, we only make inferences about the group of patients from these counties regardless of the individual demographics of these patients. Nevertheless, we plan to examine the likelihood at the individual level after accounting for the population level. In this sense, we can examine the likelihood of patients from any demographic category at the individual level accounting for the fact that these patients are from counties with increased
proportions of any demographic category. For instance, we can examine the individual likelihood of Hispanic patients from counties with increased proportions of either Black or College educational attainments.

Lastly, in this chapter, we only regarded equity as the differences in the likelihood of waiting time. However, we can also examine other factors, including the time elapsed from the moment the organ was recovered from the donor to the moment it was transplanted into the patient (i.e., ischemic time) as well as the distance traveled by organs from the donor hospital to the transplant center. Also, we can conduct multivariate analysis with multiple demographic factors. In this sense, we can simultaneously examine patients from counties with an increased proportion of both less than 10k, College, and Asian. In this chapter, our results suggest that our proposed data-intensive framework is capable of uncovering patterns of equity in organ transplantation.
Chapter 6

On the Awareness of Organ Transplantation

The current research agenda in organ transplantation is focused on increasing the number of organ donors mainly using interventions [15]. As a result, mapping the population awareness regarding organ transplantation issues is key to assess the effectiveness of these interventions. In this dissertation, we propose a sensor for capturing awareness regarding the transplantation and donation of organs using social media (Section 3.2). In this chapter, we assess how our sensor is related to donor registrations (Section 6.1), demonstrate how it can be used to characterize population awareness in the US (Section 6.2), and present a preliminary social network intervention to promote population awareness using social media (Section 6.3).
6.1 Social Media Sensor for Awareness

First, we examined the extent to which the data captured by our sensor (Section 3.2) correlates with donor registrations (Section 3.1). Indeed, the number of tweets correlates with donor registrations, exhibiting a higher correlation with most recent registration data (Figure 6.1). Precisely, organ-related tweets are more correlated with registrations from 2016 ($r = .81, p < .01$) than with registrations from 2009 ($r = .51, p < .01$) and 2012 ($r = .70, p < .01$). Interestingly, tweets are similarly correlated with the population size ($r = .86, p < .01$) regardless of the recency of population data.

![Figure 6.1: Pearson correlation among the total number of organ-related tweets, donor registrations, and population size at the state-level. Each correlation is calculated considering the number of donor registrations and population size of each year at the state-level, except for the number of organ-related tweets, for which we only have 1-year of data (Table 3.2).](image)
However, donor registrations is also correlated with the number of population, which has been increasing since 2009 ($r = .69$, $p < .01$) and becoming extremely high in 2016 ($r = .96$, $p < .01$). Precisely, a regression analysis was used to test if population significantly predicts the number of registrations, and the results indicate that population explained 48.5% of the registration variance in 2009 ($R^2 = .485$, $R^2_{adj} = .475$, $F(1, 49) = 46.2$, $p < .001$), and explained 91.7% of the registration variance in 2016 ($R^2 = .917$, $R^2_{adj} = .915$, $F(1, 49) = 542.3$, $p < .001$).

In fact, this higher correlation between donor registrations and population size is probably due to the overall increase in donor registration rates observed during this period, in which a higher proportion of the population is becoming organ donors.

To control for population size, we divide the number of tweets and registrations by the population size to obtain, respectively, the rates of tweets and registration. When we control by population size, the correlation between the tweet rates and the registration rates vanishes even in 2016 ($r = .12$, $p > .1$). Given the apparent correlation between organ-related tweets and donor registrations at the state-level, we assess this correlation at a higher resolution, focusing on the county of Los
\[ I = 8.73 \times 10^0 \]
\[ T = 3.20 \times 10^{-3} \]
\[ P = 1.01 \times 10^{-5} \]
\[ T \cdot P = 1.48 \times 10^{-8} \]

Figure 6.3: The relationship between organ-related tweets and donor registrations. (Left) A Poisson regression model of donor registrations predicted by organ-related tweets after controlling for population. The observed data is shown as solid blue dots, the model is shown as orange crosses, and the residuals are shown in the inset. (Right) Distribution of the typical number of donor registration rate for each quartile of tweet rates using 10,000 bootstrap samples.

Angeles, California, for which we have donor registration data at zip code level. However, tweets are scarce at the zip code level, and therefore we aggregated tweets and registrations at the city level (Figure 6.2). At the city level, tweet rates are correlated with registration rates \( r = .36, p < .01, 95\% \text{ CI } .12 \text{ to } .57 \). Precisely, after controlling for population size, a Poisson regression significantly predicts a 3\% increase in the number of donor registration for each extra 10 organ-related tweets (Figure 6.3, left).

We further explore the association between tweet-rates and registration rates by partitioning cities into four quartiles of tweet rates and subsequently estimating the registration rate of cities belonging to each quartile using a sufficiently high number (i.e., 10,000) of bootstrap samples (Figure 6.3). Depending on its organ-related tweet rate, a city can exhibit almost 80 additional registrations per thousand population size; while a city with a lower tweet rate typically has 202 (95\% CI,
176 to 231) registrations per thousand population size, a city with a higher rate has 280 (95% CI, 231 to 328) registrations per thousand population size.

Despite the small sample size, we also use our data-intensive framework to examine the association between tweet rates and registration rates at the city-level. Our model indicates that the tweet rate is related to donor registration (Figure 6.4, top-left); yet, this relationship might differ depending on the registration rates (Figure 6.4, top-right). Nevertheless, it also indicates a lack of statistical significance in all its estimates (Figure 6.4, bottom-left, and bottom-right), which is mainly due to the small sample size involved.

6.2 Characterization of Awareness

After evaluating our collected tweets with regards to donor registrations, in this section, we attempt to further characterize awareness in terms of organs, geographic regions, and social media users. This further characterization might allow us to uncover relationships between organs and geographic regions. Also, it might allow us to uncover groups of users who similarly mention different organs.

Organ Characterization

In the organ characterization, we attempt to characterize each organ type based on the aggregated behavior of users who tend to most mention that organ (Equation 3.1). In this sense, we can characterize all six major solid transplanted organs by presenting each of them as a histogram for which the bins are ordered according to the amount of attention given to a particular organ (Figure 6.5). For instance, livers are frequently mentioned along with kidneys, hearts, and lungs,
Figure 6.4: Data-intensive model of organ-related tweet rates and donation registration rates using our proposed data-intensive framework. (top-left) Overall, tweet rates and donor registration rates seem to be positively related. (top-right) However, the type of relationship appears to depend on the registration rate. (bottom-left) For registration rates less than or equal to the 10th percentile \( (r_{10\text{th}} = 140.3) \), a higher tweet rate seems to indicate a higher registration rate. (bottom-right) Conversely, for registration rates less than or equal to the 80\text{th} \( (r_{80\text{th}} = 383.3) \), a higher tweet rate seems to indicate a higher registration rate. However, due to a small sample size, the model shows a lack statistical significance that can be visualized as the overlapping among all confidence intervals.

respectively (Figure 6.5, top-right). As we can see, these co-occurrences are not necessarily reciprocal. Users who mostly mention hearts, livers and pancreas also tend to mention kidneys in the second place, whereas users who mostly mention intestines, kidneys, and lungs also tend to mention hearts in the second place.
Figure 6.5: Awareness characterization of the six major solid organs. Each plot shows is associated with a specific organ and shows the distribution of mentions from users who mostly mention that specific organ also mention other organs. This distribution can be seen as a row on $K$ (see equations 3.3 and 3.1). Hearts, kidneys, livers, lungs, pancreas, and intestines are depicted in red, yellow, green, blue, olive, and magenta, respectively. The histogram bars are in log scale, and the bins are ranked according to the proportion of mentions.
By focusing the characterization on most mentioned organs, we can uncover dependencies among organs, such as dual organ transplantation commonly occurring among the pairs heart-kidney, liver-kidney, and kidney-pancreas [61]. This association between organs can indicate how users who are interested in one type of organ transplantation are also interested in other types of transplantation. Interventions aiming at increasing the awareness among the population can exploit these associations. In this sense, users mostly who are interested in the transplantation of lungs might also be interested in the transplantation of hearts (Figure 6.5).

Also, this combined interest might be related to how one organ failure can lead to other organs failures [35]. For instance, individuals with heart diseases can also develop a renal dysfunction due to diabetes and hypertension [81]. Also, individuals suffering from heart diseases can develop fluid retention, which in turn damages the liver [82]. As a result, a small portion of these patients will need a liver transplant and, as the heart and the liver are damaged, their kidneys can be affected as well. Similarly, individuals suffering from liver diseases can also develop renal dysfunctions. Some of them are due to diabetes, whereas many others are due to liver disease dysfunction.

It is important to note that we are not arguing that the conversations indicate the co-occurrence of failures among different organs. Instead, we are arguing that people may have a tendency to talk about these co-occurrences. Finally, the analysis of intestine is statistically less reliable because intestine transplantation is only a small fraction of the overall organ transplants, and are related to pediatric patients [61].
Geographic Characterization

Given the correlation between geography with social and economic factors, organ transplantation awareness might not be uniformly distributed across geographic regions. In this sense, we attempt to unveil geographic patterns of awareness by exploring the characterization of regions (Equation 3.2). In this dissertation, without loss of generality, regions are US states identified using the method described in Section 3.2. Similarly to the organ characterization, a state is represented as a distribution of attention to the set of six organs (Figure 6.6). Despite significant similarities, by examining each state, we can see that every state appears to have its own distribution over the six solid organs (i.e., awareness signature).

Overall, some organs can be much more prevalent than all other organs (Figure 6.6). As a result, we are likely to find a greater number of users mentioning that organ more than all other organs. In our tweets, hearts are the most prevalent organ mentioned in almost all US states (Figure 6.6). Therefore, instead of using its prevalence, we calculate the relative risk (RR) [85] of each organ as

\[ RR_{ir} = \frac{\rho_{ir}}{\rho_{in}}, \]

where \( \rho_{ir} \) and \( \rho_{in} \) are the prevalence of mentions to organ \( i \) inside and outside the state \( r \), respectively.

In a given state, mentions to a particular organ can significantly differ from mentions to the same organ in other states. In this sense, variation in awareness can be considered as statistically significant differences in the way organs are mentioned in each state. For each organ, we use the RR to calculate the excessive incidence
Figure 6.6: Geographic awareness characterization of US states. Each state is characterized by the aggregated behavior of users inhabiting that state (Equations 3.2 and 3.3). The different shapes in the distribution of mentions to organs might indicate geographic variation in awareness. Each bin indicates the intensities of attention given to a particular organ. For instance, hearts and kidneys are the first and second-most-mentioned organ in most states, which may indicate the overall “ubiquity” of these transplants [61].

of mentions to that organ from users of a particular state relative to the overall incidence of mentions to that organ from users outside that state. Given that
Figure 6.7: Characterization of geographic variation in transplantation awareness. Each state is colored according to excessive conversations about a specific organ as measured by the relative risk \(-RR\) (see Eq. 6.1). Only statistically significant relative risks are considered as shown by the insets associated with the states of Louisiana, Massachusetts, and Rhode Island.

The distribution of \(\log(RR_{ir})\) is approximately normal [85], mentions to an organ significantly exceeds what is expected in a given state (i.e., it is highlighted), if \(\log(RR_{ir}) + z_\alpha \times \sigma_{\log(RR_{ir})} > 0\). To guarantee a 95% CI, we chose \(\alpha = 0.05\), for which \(z_\alpha = 1.96\). Most states exhibit at least one highlighted organ, for which mentions to that organ significantly exceed the national expectation (Figure 6.7).

For instance, conversations about kidneys highly exceed the expectation in Louisiana, whereas conversations about both kidneys and lungs highly exceed the expectation in Massachusetts. In the Midwestern USA, conversations about kidneys only exceed the national expectation in the state of Kansas. Interestingly, Kansas is also the only state for which a surplus of deceased kidney donors was
Figure 6.8: Awareness similarity of US states with regards to organ transplantation issues. Similarity matrix of states is visualized as a heat map, for which the lower values are associated with higher similarity. The hierarchical clustering of states is based on the distance between their awareness signatures (Figure 6.6). Using the dendrogram, we can analyze the clusters at any location in the hierarchy. From the leftmost state (i.e., Nebraska) to the rightmost state (i.e., Missouri) in the similarity matrix, the states are outlining zones of organ-related conversation in the following order: livers (from Delaware to North Dakota), lungs (from Massachusetts to Wisconsin), kidneys (from New York to Virginia) and hearts (from Minnesota to California). Similarly, states without a highlighted organ tend to cluster, for instance, in the zone between New Mexico and Indiana.

These associations are generally discovered because health-related traits are geographically correlated. For instance, a higher risk of hypertension mainly due to diet is consistently observed in the Stroke Belt, a commonly known region in
the Southern USA. Similarly, an increased incidence of liver diseases probably associated with both diet and genetic traits are observed in the Western USA.

Besides uncovering the highlighted organs at each state due to their significantly excess of conversations, we also explore the similarities among states by examining the way different organs are consistently mentioned by users from these states. To investigate this similarity, we use the agglomerative clustering algorithm [69] because of its simplicity as well as its flexibility, allowing us to examine the resultant clusters at different levels on the hierarchy. In the clustering, each state is a data-point (rows of matrix $K$), and the feature vector is the distribution of mention to each organ from users of that state (Figure 6.6). Lastly, we use the Hellinger distance as our distance metric because it is more suitable for comparing discrete probability distributions and, in particular, distributions that are highly skewed [37].

The clustering results suggest that states tend to group into clusters depending on the similarity in which their users mention different organs (Figure 6.8). These clusters exhibit some consistency with the aforementioned organs highlighted at each state such that states exhibiting the same highlighted organ tend to be grouped together (Figure 6.7). For instance, livers are highlighted in the states of Delaware, Rhode Island, and Colorado, whereas lungs are highlighted in the of states of Oregon, Georgia, and Virginia.

**User Characterization**

The aforementioned characterizations of organs and states were capable of capturing patterns of awareness hidden in organ-related tweets. These characterizations were based on the aggregated behavior of users who were
partitioned based on the organ to which they give their maximum attention or the specific state in which they inhabit. Besides partitioning users, we also attempt to characterize their full behavior, grouping them into clusters of users who tend to mention different organs in similar ways.

To cluster the users, we use the algorithm $k$-means with the number of clusters $k$ equals to 12. After some empirical evaluation, this number of clusters was associated with convenient values of inertia (i.e., guaranteeing convergence), average cluster size (i.e., avoiding finding clusters with small size), and silhouette coefficient (i.e., finding appropriate clusters given the data). As expected, users tend to be grouped into different clusters depending on the way they mention different organs (Figure 6.9). Given that we are characterizing six different organs, the number of clusters $k$ could plausibly assume values greater than six, allowing at least one cluster for each organ. A greater number of clusters could still be used because even a cluster associated with 0.3% of users is still related to roughly 2 thousand users.

Certainly, these clusters deserve further exploration. Nevertheless, they appear to be capable of revealing information that has the potential to be used to identify general patterns exhibited by different classes of users. Specifically, we can identify four groups of users who mostly tend to mention a single organ (Figure 6.9, top). For instance, some users tend to mostly mention hearts (in red), whereas some users tend to mostly mention kidneys (in orange).

Similarly, we can also identify users who tend to mention two and even three different organs. Some users appear to mention virtually all organs, especially when compared with other clusters (Figure 6.9, bottom-rightmost). Possibly, these clusters might be associated with organ-related users with different
Figure 6.9: Groups of users who mention different organs in a similar way. To cluster the users, we used the algorithm $k$-means with the number of clusters $k$ equals to 12. This number of clusters was empirically determined and was associated with silhouette coefficient, average cluster size and inertia 0.953, 31697.42 and 2512.27, respectively. For each cluster, we show the distribution of mentions to different organs and its relative size. Possibly, these clusters might be associated with users who play different roles in the context of organ donation.

attitudes towards organ donation, and this information can be used along with the geographic characterization, for instance, to investigate possible correlations further. In the next section, we report on a preliminary social network intervention using social media in which we had the opportunity to collaborate to increase the awareness towards organ donation among populations from minority groups.
6.3 Social Network Intervention

Despite considerable efforts attempting to increase the number of organ donors over the past 10 years, only half of US adults are currently registered as organ donors\(^1\). Given the ever-increasing incidence of end-stage organ failure, the shortage of organs has been deepening, and we are currently witnessing a greater patient need for organ transplantation [93]. Particularly, African American and Hispanic populations are at a disadvantage due to the scarcity of registered donors with appropriate levels of histocompatibility [83, 14]. This lack of organ donors from underrepresented demographics is mainly due to insufficient health education among these populations [14].

To tackle this issue, we have collaborated in the development of a Social Network Intervention (SNI) aimed at improving awareness about organ donation among minorities by motivating them to engage with the organ registration website\(^2\). This SNI was a preliminary intervention using social media and was conducted by our collaborators from the Department of Medicine at the University of California, Los Angeles (UCLA), who kindly provided us all the data collected during the intervention. Our role in this intervention was limited to the data analysis of the collected data during the intervention, which consisted in the number of impressions, clicks, and page views (Tables C.1 and C.2). Nevertheless, to the best of our knowledge, we will attempt to describe the intervention inasmuch detail as possible, including its motivation and intended audience.

\(^1\)Donate Life 2017 Annual Report available at https://www.donatelifeline.net

\(^2\)Donate Life California, Sacramento, CA https://register.donatelifecalifornia.org/register/
Figure 6.10: Effectiveness measures collected during our social network intervention (SNI). The measures are the (A) Impressions, Clicks, and Page Views and their normalized versions (B) Clicks per Impression and Page Views per Impression. See Tables C.1 and C.2 for details.

The SNI was implemented from August 4th, 2016 to September 3rd, 2016 within the greater Los Angeles area, and targeted individuals who self-identified as Hispanics, from lower income levels, with age from 18 to 65 years, and with at least some high school education (Figure C.1). By using the advertisement platform
provided by Facebook, the SNI consisted of delivering to the tailored audience short educational content regarding organ donation. From August 4th to August 23rd, using a third-party optimization software\(^3\), the intervention identified among the educational contents, the one with the highest capability to increase engagement among the audience. From August 24th, the intervention only delivered the optimized content, playing a key role in exposing the most appealing content to the targeted audience. To the best of our knowledge, the audience targeted after the optimization has similar demographics to the audience before the optimization.

We used the impressions \((I)\), clicks \((C)\), and page views \((P)\) collected during the intervention to measure the effectiveness of the social network intervention (Figure 6.10(A)). Given that we had no information prior to the intervention to use as the baseline, we considered the optimization as an instrumental variable and used the collected data before the optimization as a control group. Since the population in the control group is also receiving the intervention, the control group can be considered as a conservative control group.

As expected, the number of clicks and page views positively correlate with the number of impressions \((r = .84, p < .001)\), and this correlation increases \((r = .94, p < .001)\) after the optimization (Figure 6.11). Precisely, after the optimization, the number of clicks increases 10 times, from 503 to 3,488, while the number of page views increases 7 times, from 185 to 1,287. Given that the number of impressions was three times higher after the optimization, increasing from 18,630 to 54,290 impressions (Tables C.1 and C.2), we normalize the number of clicks

\(^3\)MAV12 Inc., Santa Monica, CA
Figure 6.11: Correlation between effectiveness measures. The optimization plays a key role in affecting the Impressions, Clicks, and Page Views as well as their normalized versions Clicks per Impression and Page Views per Impression. Overall, these metrics become more positively associated with the optimization. 

\[(C/I)\) and page views \((P/I)\) by the number of impressions to account for different levels of resource utilization before and after the optimization (Figure 6.10(B)).

Although the number of clicks per impression negatively correlates with the number of impressions before the optimization \((r = -.67, p < .005)\), this negative correlation vanishes \((r = -.48, p > .10)\) after the optimization (Figure 6.11). Using the optimized content after the optimization, additional 21 (95% CI, 8 to 35) clicks can be overall obtained per thousand of impressions, with the number of clicks per thousand impressions increasing from 42 (95% CI, 35 to 48) to 63 (95% CI, 50 to 77) (Figure 6.12(A)). However, the number of clicks per impression can saturate with the number of impressions thereby making the intervention less efficient as the number of impressions increases. Indeed, before the optimization, as the number of impressions increases, the number of clicks per thousand impressions decreases from 41 (95% CI, 40 to 41) to 21 (95% CI, 10 to 31). Interestingly, this saturation
vanishes after the optimization, and the number of clicks per impression is not statistically affected as the number of impressions increases.

Similarly, the number of page views per impression negatively correlates with the number of impressions \( r = -0.73, p < 0.05 \) before the optimization, but this negative correlation vanishes \( r = 0.51, p > 0.10 \) after the optimization (Figure 6.11). Before the optimization, as the number of impressions increases, the number of page views per thousand impression decreases from 21 (95% CI, 21 to 22) to 5 (95% CI, −3 to 12) (Figure 6.12(B)). Although not statistically significant, after the optimization, at least additional 3 page visits per thousand of impressions has the potential to be obtained as the number of impressions increases.
By considering clicks and page views as positive attitudes towards organ donation, we show that targeting a focused audience with a tailored content is key to make the intervention more effective, converting more impressions into attitudes towards organ donation registration. One key aspect of the SNI is its capability to organically increase the number of impressions by users who share with their social ties the educational content to which they were exposed. Our results indicate that the SNI is capable of reaching and engaging the audience, potentially increasing awareness that can lead to an increase in the number of organ donors.

6.4 Discussions

The population awareness towards organ transplantation issues is particularly relevant to increase the number of organ donors. In this chapter, we demonstrated how social media has the potential to be used as a sensor to characterize the population awareness towards organ transplantation issues as well as its potential to increase awareness towards organ donation, particularly among populations from minority groups. In this sense, the sensor we develop in this dissertation can provide a real-time characterization of awareness that has the potential to not only assess the effectiveness of social network interventions such as the one aforementioned, but also to support in the design of social network interventions that are more tailored to the cultural, social, religious, and educational differences between regions.

Nevertheless, the sensor has some limitations mainly regarding the bias in the collected data. For instance, Twitter users are not a uniform sample of the US population especially with regards to geography, gender and ethnicity/race [51].
Specifically, Twitter users are typically biased towards highly populated counties and male users. Also, different ethnicity/race (i.e., Caucasian, African-American, Asian, and Hispanic) can be over or under-sampled depending on the region. For instance, the Midwestern population of United States is underrepresented among Twitter users.
Chapter 7

Conclusions, Limitations, and Future Works

In this dissertation, we proposed a comprehensive data-driven characterization of organ transplantation with regards to the aspects of efficiency, equity, and awareness. To achieve this comprehensive characterization, we created a sensor for population awareness using social media and proposed a nonparametric probabilistic data-intensive framework. Lastly, we characterized efficiency, equity, and awareness using our proposed data-intensive framework as well as common approaches in statistics and data science.

In the context of organ transplantation, our sensor contributes to the lack of data on the population awareness about organ transplantation issues. Also, it allows for real-time characterization of the population awareness. This real-time characterization is of particular relevance to support interventions aiming at increasing the number of organ donors. In this sense, this sensor has the potential to both inform interventions as well as assess their effectiveness.
Similarly, in the context of organ transplantation, our data science approach contributes to the lack of a unified framework to characterize efficiency, equity, and awareness in organ transplantation using a framework that is nonparametric, probabilistic, and data-intensive. Currently, we suggest that our proposed data-intensive framework can complement existing methodology in the assessment of organ transplantation. Besides, our data-intensive framework can also be extended to characterize organ transplantation in countries other than the United States as well as to study aspects of health other than organ transplantation (e.g., cardiovascular disease). Moreover, our framework is not necessarily limited to the context of health and medical sciences; it has the potential to be generalized to the study of phenomena in the social (e.g., patterns of crime) and physical (e.g., patterns of particles movements) sciences.

In the characterization of efficiency, we defined efficiency as the probability of graft failure after transplantation. Then, using our data-intensive framework, we demonstrated how this probability changes depending on three factors: the distance traveled by organs from the donor hospital to the transplant center, the ischemic time elapsed from the moment organs are recovered from donors and transplanted into patients, and the waiting time elapsed since the patients enter in the waiting list to the moment they receive organ transplants. For each factor, we separately characterized the efficiency of different organs and demonstrated different patterns of efficiency. In the characterization of equity, we defined equity as the change in the probability of waiting for organ transplants. Using our data-intensive framework, we demonstrated how this probability changes depending on three groups of demographic factors: race/ethnicity, educational attainment, and income.
Our proposed data-driven characterization also is general enough to have the potential to be instantiated for countries other than the United States. In the case of an English speaking country (e.g., United Kingdom), we will only need the transplantation as well as demographic data for that country. In the case of a non-English speaking country (e.g., Brazil), we need to configure our sensor to also capture tweets in languages other than English (e.g., Portuguese).

Limitations

In this dissertation, we demonstrated that our nonparametric probabilistic data-intensive framework is capable of uncovering patterns from the data. However, we neglected to examine the minimum necessary data required for the data-intensive framework to yield statistically significant estimates. Also, we also neglected to examine how the optimally decide on the resolution of the data binning strategy as determined by the number of bins. This resolution is ultimately limited to the available computational power. Although our framework is highly scalable, excessively higher resolutions might only unnecessarily increase the computational cost.

In the characterization of efficiency, we provided the results for all six organs. Yet, in the characterizations of efficiency and equity, we mainly demonstrated the ability of our proposed data-intensive framework to uncover patterns for hearts and kidneys. As a result, we still cannot generalize its capability of uncovering patterns for all six solid organs. In particular, our data-intensive framework appears to be sensitive the significantly small sample sizes. For instance, our framework was not able to provide statistically significant estimates to the patterns uncovered in the
characterization of awareness. Although we have not analyzed the results, they suggest that our data-intensive framework is capable of uncovering patterns for all types of organ transplants. However, as we previously mentioned, we still need to examine the minimum necessary data for our framework to yield statistically significant results.

In the characterization of equity, our results should be interpreted taking into account the correlation among demographic factors. In this sense, we attempted to demonstrate the correlation between demographic factors using our data-intensive framework. Despite its flexibility to account for this correlation, our data-intensive framework still lacks a methodology to take this correlation into account.

In the characterization of awareness, our sensor collects tweets containing words that are supposedly related to organ transplantation issues. Indeed, our results suggest that our collected tweets are associated with organ donation registrations. Also, we manually validated that the content of a few randomly selected tweets is related to organ transplantation issues. However, our methodology still lacks a robust validation regarding the content of tweets using the overall data in an automatic manner.

Also, our sensor is sensitive to the language used to collect the tweets. In this dissertation, we limited the words to the English language. Given the significant number of non-English speakers in the United States, especially in specific counties and zip codes, our sensor might underestimate the number of organ-related tweets in these regions. Also, our sensor is sensitive to the population demographics. The users on Twitter are not a uniform sample of the population. As a result, our collected tweets are likely misrepresenting the actual population depending on their geographic regions.
Future works

Our results indicate that our nonparametric probabilistic data-intensive framework has the potential to be generalized to uncover patterns from any data set. In this sense, in future works, we want to assess this generalization capability of our data-intensive framework. Before, we still need to examine what is the minimally necessary data required by our data-intensive framework as well as how to optimally set the data binning resolutions. Then, we will validate the methodology for organ transplantation across different organs as well as using transplantation data from other countries, including developing countries. Lastly, we will extend our data-intensive framework using data sets related to other subjects in health (e.g., cardiovascular disease), social sciences (e.g., crime), and physical sciences (the movement of particles).

We also plan to further work on our sensor for population awareness and its generalization capability. Before, will also examine the minimal amount of data required by the sensor to yield statistically significant results. Then, we will include languages other the English and include subjects other than organ transplantation. Lastly, we will further explore the distinct classes of users associated with different patterns of awareness.

By using our data-intensive framework, we have the potential to characterize each individual aspects as well as the trade-off between then in a unified manner. For instance, we examined how waiting time affects the probability of graft failure, how the race/ethnicity affects the probability of waiting time, and how the number of organ-related tweets affects the probability of donation registrations. In the characterization of equity, we demonstrated how changes in the likelihood of
waiting time seem to affect the likelihood of graft failure. However, we will further examine the trade-off between these relationships.
Bibliography


Evaluating the Healthcare System: Effectiveness, Efficiency, and Equity. Health 


progress and the continuing problem. Journal of Biomedical Research, pages 
81–82, mar 2016.

Salvalaggio, Dorry L. Segev, Sommer E. Gentry, Janet Tuttle-Newhall, and 
Krista L. Lentine. The interplay of socioeconomic status, distance to center, 
and interdonor service area travel on kidney transplant access and outcomes. 

allocation for liver transplantation: Tackling geographic disparity through 


Appendix A

Benchmarking of Data-Intensive Model

In this section, we provide the results of our proposed data-intensive nonparametric probabilistic approach with regards to the following well-known benchmarking functions:

1. Negative correlation (Figure A.1, left column): a sample of 10,000 data points from a Multivariate Gaussian $\mathcal{N}[\mu, \Sigma]$, with $\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, and $\Sigma = \begin{bmatrix} 1 & -.6 \\ -.6 & 1 \end{bmatrix}$;
2. Hidden correlation (Figure A.1, right column): a sample of 10,000 data points from a Multivariate Gaussian $\mathcal{N}[\mu, \Sigma]$, with $\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, and $\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ and a sample of 1,000 from a Multivariate Gaussian $\mathcal{N}[\mu, \Sigma]$, with $\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, and $\Sigma = \begin{bmatrix} 1 & .6 \\ .6 & 1 \end{bmatrix}$;

3. Hidden sinusoidal function (Figure A.2, left column): a sample of 10,000 data points from a from a Multivariate Gaussian $\mathcal{N}[\mu, \Sigma]$, with $\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, and $\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ and a sample of 10,000 from a sinusoidal function $f(x) = \sin(2\pi x)$, with $X$ uniformly distributed from $-1$ to 1;

4. Simpson’s paradox (Figure A.2, right column): a sample of 10,000 data points from a Multivariate Gaussian $\mathcal{N}[\mu, \Sigma]$, with $\mu = \begin{bmatrix} 3 \\ 8 \end{bmatrix}$, and $\Sigma = \begin{bmatrix} 1 & .6 \\ .6 & 1 \end{bmatrix}$ and a sample of 10,000 data points from a Multivariate Gaussian $\mathcal{N}[\mu, \Sigma]$, with $\mu = \begin{bmatrix} 10 \\ 3 \end{bmatrix}$, and $\Sigma = \begin{bmatrix} 1 & .6 \\ .6 & 1 \end{bmatrix}$;
One can see our data-intensive model uncover a hidden relationship between $X$ and $Y$ (Figure A.1, left). As $X$ ranges from $-4$ to $4$, the probability of values $Y$ less than or equal to $Y$ slightly decreases. Similarly, the data-intensive models unveils a sinusoidal function that is hidden among a uncorrelated data (Figure A.2, left). The change in probability directly arrives from the existence of statistically significant mass of data-points $(X,Y)$ associated according to a sinusoidal relationship that was artificially generated. Lastly, the data-intensive models uncovers the existence of two regimes of statistically associated data as $X$ moves from $0$ to $12$ (Figure A.2, right).
Figure A.1: Benchmarking of our data-intensive approach using a negative correlation and a hidden correlation. (Left Column) A sample of 10,000 data points from a Multivariate Gaussian $\mathcal{N}(\mu, \Sigma)$, with $\mu_X = \mu_Y = 0$, $\sigma_X = \sigma_Y = 1$, and $\sigma_{XY} = \sigma_{YX} = -0.6$. (Right Column) A sample of 10,000 data points from a Multivariate Gaussian $\mathcal{N}(\mu, \Sigma)$, with $\mu_X = \mu_Y = 0$, $\sigma_X = \sigma_Y = 1$, and $\sigma_{XY} = \sigma_{YX} = 0$, as well as a sample of 1,000 from a Multivariate Gaussian $\mathcal{N}(\mu, \Sigma)$, with $\mu_X = \mu_Y = 0$, $\sigma_X = \sigma_Y = 1$, and $\sigma_{XY} = \sigma_{YX} = 0.6$. Our data-intensive model uses a $E$ data binning with $n$ equals to 30 and $max$ equals to 99th percentile for both $X$ and $Y$; the min values of $X$ and $Y$ are set to 0th and 10th percentiles, respectively.
Figure A.2: Benchmarking of our data-intensive approach using a hidden sinusoidal function and the Simpson’s paradox. (Left Column) A sample of 10,000 data points from a Multivariate Gaussian $\mathcal{N}(\mu, \Sigma)$, with $\mu_X = \mu_Y = 0$, $\sigma_X = \sigma_Y = 1$, and $\sigma_{XY} = \sigma_{YX} = 0$, as well as a sample of 10,000 from a sinusoidal function $f(x) = \sin(2\pi x)$, with $X$ uniformly distributed from $-1$ to $1$. (Right Column) A sample of 10,000 data points from a Multivariate Gaussian $\mathcal{N}(\mu, \Sigma)$, with $\mu_X = 10$, $\mu_Y = 3$, $\sigma_X = \sigma_Y = 1$, and $\sigma_{XY} = \sigma_{YX} = .6$. Our data-intensive model uses a $E$ data binning with $n$ equals to 30 and $\max$ equals to 99th percentile for both $X$ and $Y$; the $\min$ values of $X$ and $Y$ are set to 0th and 10th percentiles, respectively.
Appendix B

On the Efficiency of Organ Transplantation

In Chapter 4, we provide the results of our proposed nonparametric probabilistic approach for hearts and kidneys. In this appendix, we provide the results for all six major solid organs. For the factor distance, the raw data, the heatmaps, and the curves are shown in figures B.1, B.2, and B.3, respectively. For the factor ischemic time, the raw data, the heatmaps, and the curves are shown in figures B.4, B.5, and B.6, respectively. Lastly, for the factor waiting time, the raw data, the heatmaps, and the curves are shown in figures B.7, B.8, and B.9, respectively.
Figure B.1: Raw data visualization of the association of graft failure with distance for kidneys, hearts, livers, lungs, pancreas, and intestines. It is important to note that the data visualization is employing a level of transparency to allow us distinguish between the locations where there is a high incidence of data points from the locations where the data is sparse.
Figure B.2: Association of early graft failures and distance for kidneys, hearts, livers, lungs, pancreas, and intestines using our nonparametric probabilist approach (3.3). The distance $D$ (x-axis) ranges from 0 to 99\textsuperscript{th} distance percentile $d_{99}$ for each organ using 10 miles bins. The graft failure time $T$ in days (y-axis) ranges from 0 (i.e., grafts that failed in the same day they were transplanted) to 30 days. Data visualization in terms of a heat map, where each cell represents $Pr(T \leq t | D > d)$ as estimated by Equation 3.5.
Figure B.3: Association of early graft failures and distance for kidneys, hearts, livers, lungs, pancreas, and intestines using our nonparametric probabilist approach (3.3). The distance $D$ (x-axis) ranges from 0 to 99th distance percentile $d_{99\text{th}}$ for each organ using 10 miles bins. The graft failure time $T$ in days (y-axis) ranges from 0 (i.e., grafts that failed in the same day they were transplanted) to 30 days. Data visualization in terms of curves, where the y-axis represents $\Pr(T \leq t|D > d)$ for 5 different values of $t \{0, 1, 7, 15, 30\}$. 
Figure B.4: Raw data visualization of the association of graft failure with ischemic time for kidneys, hearts, livers, lungs, pancreas, and intestines. It is important to note that the data visualization is employing a level of transparency to allow us to distinguish between the locations where there is a high incidence of data points from the locations where the data is sparse.
Figure B.5: Association of early graft failures and ischemic time for kidneys, hearts, livers, lungs, pancreas, and intestines using our nonparametric probilist approach (3.3). The ischemic time $I$ in 1 hour bins (x-axis) ranges from 0 to the 99th ischemic time percentile of each organ $i_{99th}$. The graft failure time $T$ in days (y-axis) ranges from 0 (i.e., grafts that failed in the same day they were transplanted) to 30 days. Data visualization of the nonparametric probabilist approach in terms of a heat map, where each cell represents $\Pr(T \leq t | I > i)$ as estimated by Equation 3.5.
Figure B.6: Association of early graft failures and ischemic time for kidneys, hearts, livers, lungs, pancreas, and intestines using our nonparametric probabilist approach (3.3). The ischemic time $I$ in 1 hour bins (x-axis) ranges from 0 to the 99th ischemic time percentile of each organ $i_{99th}$. The graft failure time $T$ in days (y-axis) ranges from 0 (i.e., grafts that failed in the same day they were transplanted) to 30 days. Data visualization of the nonparametric probabilistic approach in terms of curves, where the y-axis represents $\Pr(T \leq t | I > i)$ for 5 different values of $t \{0, 1, 7, 15, 30\}$. 
Figure B.7: Raw data visualization of the association of graft failure with waiting time for kidneys, hearts, livers, lungs, pancreas, and intestines. It is important to note that the data visualization is employing a level of transparency to allow us to distinguish between the locations where there is a high incidence of data points from the locations where the data is sparse.
Figure B.8: Association of early graft failures and waiting time for kidneys, hearts, livers, lungs, pancreas, and intestines using our nonparametric probabilist approach (3.3). The waiting time $W$ in 1 hour bins (x-axis) ranges from 0 to the 99th waiting time percentile of each organ $i_{99th}$. The graft failure time $T$ in days (y-axis) ranges from 0 (i.e., grafts that failed in the same day they were transplanted) to 30 days. Data visualization of the nonparametric probabilist approach in terms of a heat map, where each cell represents $\Pr(T \leq t | W > w)$ as estimated by Equation 3.5.
Figure B.9: Association of early graft failures and waiting time for kidneys, hearts, livers, lungs, pancreas, and intestines using our nonparametric probabilist approach (3.3). The waiting time $W$ in 1 hour bins (x-axis) ranges from 0 to the 99th waiting time percentile of each organ $i_{99th}$. The graft failure time $T$ in days (y-axis) ranges from 0 (i.e., grafts that failed in the same day they were transplanted) to 30 days. Data visualization of the nonparametric probabilistic approach in terms of curves, where the y-axis represents $Pr(T \leq t | W > w)$ for 5 different values of $t \{0, 1, 7, 15, 30\}$.  

157
Appendix C

Social Network Intervention

In this Appendix, we provided additional information regarding the social network intervention (SNI) in which we had the opportunity to collaborate. The first information is the demographics of the targeted audience. Overall, we matched the targeted demographics in terms of economic status as measured by household income as well as household ownership (Figure C.1). However, the SNI was not effective in reaching younger age groups, and was neither effective in reaching men in the same quantity as women.

The SNI was implemented from August 4th 2016 to September 3rd 2016. The collected data consist of a set of metrics associated with the engagement of organ donation intervention before (Table C.1) and after (Table C.2) the optimization was implemented. This set of metrics includes the number of impressions ($I$), clicks ($C$), and page views ($P$). To assess the effectiveness of the intervention, we model the number of clicks per impression and page views per impressions based on the number of impressions as well as based on whether the optimization was used (Table C.3).
Figure C.1: Demographics of the audience targeted by our Social Network Intervention (SNI) provided by Facebook. The audience was targeted based on gender, age, household ownership, and household income. Each plot shows the targeted and matched population demographics. For instance, the targeted population was successfully matched in terms of household ownership.
Table C.1: Social Network Intervention (SNI) Engagement Effectiveness before Optimization. From Aug 4th to Sept 3rd, we daily obtained three measures provided by Facebook: the number of impressions ($I$), the number of clicks ($C$), and the number of page views ($P$). Clicks per impression ($C/I$) and page views per impression ($P/I$) are normalized versions obtained by dividing the number of clicks and page views by the number of impressions. From August 4th to August 23rd, the intervention identified among the educational contents, the one with the highest capability to increase engagement among the audience.

<table>
<thead>
<tr>
<th></th>
<th>$I$</th>
<th>$C$</th>
<th>$P$</th>
<th>$C/I$ (%)</th>
<th>$P/I$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aug-04</td>
<td>4,639</td>
<td>198</td>
<td>102</td>
<td>4.27%</td>
<td>2.20%</td>
</tr>
<tr>
<td>Aug-05</td>
<td>8,831</td>
<td>346</td>
<td>200</td>
<td>3.92%</td>
<td>2.26%</td>
</tr>
<tr>
<td>Aug-06</td>
<td>11,058</td>
<td>412</td>
<td>204</td>
<td>3.73%</td>
<td>1.84%</td>
</tr>
<tr>
<td>Aug-07</td>
<td>14,731</td>
<td>544</td>
<td>290</td>
<td>3.69%</td>
<td>1.97%</td>
</tr>
<tr>
<td>Aug-08</td>
<td>24,697</td>
<td>699</td>
<td>272</td>
<td>2.83%</td>
<td>1.10%</td>
</tr>
<tr>
<td>Aug-09</td>
<td>28,165</td>
<td>563</td>
<td>237</td>
<td>2.00%</td>
<td>0.84%</td>
</tr>
<tr>
<td>Aug-10</td>
<td>31,336</td>
<td>778</td>
<td>242</td>
<td>2.48%</td>
<td>0.77%</td>
</tr>
<tr>
<td>Aug-11</td>
<td>23,904</td>
<td>602</td>
<td>172</td>
<td>2.52%</td>
<td>0.72%</td>
</tr>
<tr>
<td>Aug-12</td>
<td>21,661</td>
<td>578</td>
<td>172</td>
<td>2.67%</td>
<td>0.79%</td>
</tr>
<tr>
<td>Aug-13</td>
<td>17,584</td>
<td>501</td>
<td>167</td>
<td>2.85%</td>
<td>0.95%</td>
</tr>
<tr>
<td>Aug-14</td>
<td>16,884</td>
<td>417</td>
<td>124</td>
<td>2.47%</td>
<td>0.73%</td>
</tr>
<tr>
<td>Aug-15</td>
<td>22,518</td>
<td>585</td>
<td>198</td>
<td>2.60%</td>
<td>0.88%</td>
</tr>
<tr>
<td>Aug-16</td>
<td>20,854</td>
<td>523</td>
<td>188</td>
<td>2.51%</td>
<td>0.90%</td>
</tr>
<tr>
<td>Aug-17</td>
<td>19,964</td>
<td>458</td>
<td>168</td>
<td>2.29%</td>
<td>0.84%</td>
</tr>
<tr>
<td>Aug-18</td>
<td>18,252</td>
<td>435</td>
<td>161</td>
<td>2.38%</td>
<td>0.88%</td>
</tr>
<tr>
<td>Aug-19</td>
<td>15,264</td>
<td>353</td>
<td>126</td>
<td>2.31%</td>
<td>0.83%</td>
</tr>
<tr>
<td>Aug-20</td>
<td>16,552</td>
<td>381</td>
<td>168</td>
<td>2.30%</td>
<td>1.01%</td>
</tr>
<tr>
<td>Aug-21</td>
<td>17,594</td>
<td>392</td>
<td>148</td>
<td>2.23%</td>
<td>0.84%</td>
</tr>
<tr>
<td>Aug-22</td>
<td>15,061</td>
<td>528</td>
<td>138</td>
<td>3.51%</td>
<td>0.92%</td>
</tr>
<tr>
<td>Aug-23</td>
<td>22,975</td>
<td>784</td>
<td>228</td>
<td>3.41%</td>
<td>0.99%</td>
</tr>
</tbody>
</table>
Table C.2: Social Network Intervention (SNI) Engagement Effectiveness after Optimization. From Aug 4 to Sept 3, we obtained three measures provided by Facebook daily: the number of impressions ($I$), the number of clicks ($C$), and the number of page views ($P$). Clicks per impression ($C/I$) and page views per impression ($P/I$) are normalized versions obtained by dividing the number of clicks and page views by the number of impressions. From August 24th, the intervention only delivered the optimized content, playing a key role in exposing the most appealing content to the targeted audience.

<table>
<thead>
<tr>
<th>Social Network Intervention After Optimization</th>
<th>I</th>
<th>C</th>
<th>P</th>
<th>C/I (%)</th>
<th>P/I (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aug-24</td>
<td>53,280</td>
<td>2,708</td>
<td>825</td>
<td>5.08%</td>
<td>1.55%</td>
</tr>
<tr>
<td>Aug-25</td>
<td>54,076</td>
<td>3,154</td>
<td>1,007</td>
<td>5.83%</td>
<td>1.86%</td>
</tr>
<tr>
<td>Aug-26</td>
<td>47,259</td>
<td>2,778</td>
<td>819</td>
<td>5.88%</td>
<td>1.73%</td>
</tr>
<tr>
<td>Aug-27</td>
<td>55,165</td>
<td>3,067</td>
<td>898</td>
<td>5.56%</td>
<td>1.63%</td>
</tr>
<tr>
<td>Aug-28</td>
<td>67,832</td>
<td>3,882</td>
<td>1,485</td>
<td>5.72%</td>
<td>2.19%</td>
</tr>
<tr>
<td>Aug-29</td>
<td>72,089</td>
<td>4,243</td>
<td>1,664</td>
<td>5.89%</td>
<td>2.31%</td>
</tr>
<tr>
<td>Aug-30</td>
<td>79,789</td>
<td>4,679</td>
<td>1,721</td>
<td>5.86%</td>
<td>2.16%</td>
</tr>
<tr>
<td>Aug-31</td>
<td>88,074</td>
<td>4,967</td>
<td>2,041</td>
<td>5.64%</td>
<td>2.32%</td>
</tr>
<tr>
<td>Sep-01</td>
<td>96,850</td>
<td>5,118</td>
<td>2,118</td>
<td>5.28%</td>
<td>2.19%</td>
</tr>
<tr>
<td>Sep-02</td>
<td>88,455</td>
<td>4,770</td>
<td>1,975</td>
<td>5.39%</td>
<td>2.23%</td>
</tr>
<tr>
<td>Sep-03</td>
<td>99,190</td>
<td>4,545</td>
<td>1,643</td>
<td>4.58%</td>
<td>1.66%</td>
</tr>
</tbody>
</table>

Table C.3: Effectiveness of the social network intervention using Poisson regression models. We model clicks per impression ($C/I$) and page views per impression ($P/I$) based on whether the optimization is used ($O$) as well as based on the number of impressions ($I$). We report the estimation of each coefficient, its standard error, and its significance level.

<table>
<thead>
<tr>
<th>C/I</th>
<th>P/I</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept 4.1E-02*** (3.3E-03)</td>
<td>2.2E-02*** (2.3E-03)</td>
</tr>
<tr>
<td>$O$  2.1E-02*** (6.7E-03)</td>
<td>-8.1E-03 (4.8E-03)</td>
</tr>
<tr>
<td>$I$  5.9E-07*** (1.7E-07)</td>
<td>-5.8E-07*** (1.2E-07)</td>
</tr>
<tr>
<td>$O·I$ 5.9E-07*** (1.8E-07)</td>
<td>6.6E-07*** (1.3E-07)</td>
</tr>
<tr>
<td>$R^2$ 0.90</td>
<td>0.74</td>
</tr>
</tbody>
</table>

Standard errors in parentheses. * $p \leq .1$, ** $p \leq .05$, ***$p \leq .01$