An Analytics Approach to Problems in Health Care

by

Jerry Lai Kung

A.B., Harvard College (2011)
M.A.St., University of Cambridge (2012)

Submitted to the Sloan School of Management
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Operations Research

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2017

© Massachusetts Institute of Technology 2017. All rights reserved.

Author .................................................................
Sloan School of Management
May 15, 2017

Certified by ..............................................................
Dimitris Bertsimas
Boeing Professor of Operations Research
Co-director, Operations Research Center
Thesis Supervisor

Accepted by ............................................................
Patrick Jaillet
Dugald C. Jackson Professor
Department of Electrical Engineering and Computer Science
Co-Director, Operations Research Center
An Analytics Approach to Problems in Health Care

by

Jerry Lai Kung

Submitted to the Sloan School of Management on May 15, 2017, in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Operations Research

Abstract

Health care expenditures in the United States have been increasing at unsustainable rates for more than thirty years with no signs of abating. Decisions to accept or reject deceased-donor kidneys offered to patients on the kidney transplantation waitlist currently rely on physician experience and intuition. Scoring rules to determine which end-stage liver disease patients are in most dire need of immediate transplantation have been haphazardly designed and reactively modified in an attempt to decrease waitlist mortality and increase fairness for cancer patients.

For each of the above problem settings, we propose a framework that takes real-world data as input and draws upon modern data analytics methods ranging from mixed integer linear optimization to predictive machine learning to yield actionable insights that can add a significant edge over current practice.

We describe an approach that, given insurance claims data, leads conservatively to a 10% reduction in health care costs in a study involving a large private US employer. Using historical data for patients on the kidney waitlist and organ match runs, we build a model that achieves an out-of-sample AUC of 0.87 when predicting whether or not a patient will receive a kidney of a particular quality within three, six, or twelve months. Given historical data for patients on the liver waitlist, we create a unified model that is capable of averting an additional 25% of adverse events in simulation compared to current practice without disadvantaging cancer patients.

Thesis Supervisor: Dimitris Bertsimas
Title: Boeing Professor of Operations Research
Co-director, Operations Research Center
Acknowledgments

I wish to acknowledge and thank my advisor, Dimitris Bertsimas, for all of his thoughtful and kind-hearted support over the last five years. He has been my number one advocate throughout my time at MIT, and it has been a tremendous privilege to be mentored by an individual with an unparalleled clarity of research vision and a singular devotion to improving the human condition through such novel research.

I would also like to thank the two other members of my thesis committee: Nikolaos Trichakis and Vivek Farias. In particular, Nikos has been extremely helpful in taking the time to lend valuable insights that were indispensable in the preparation of the second and third parts of this thesis. I am grateful to Vivek for also serving on my general examination committee and providing excellent feedback throughout.

I thank Laura Rose and Andrew Carvalho for their tireless efforts in making the ORC run smoothly day-to-day. I am also very thankful for the assistance of Stephen Sofoul and Benefits Science Technologies for making the data available for the first part of this thesis. I thank Parsia Vagefi and David Wojciechowski for all of the fruitful discussions regarding kidneys and livers.

I am especially grateful for all of the students at the ORC, of which there are too many to name, with whom I have had the immense pleasure of learning, collaborating, and socializing. Special mention goes to Martin Copenhaver and Iain Dunning for the many existential discussions over chicken and memes.

I thank David Parkes for his continued mentorship. My college friends have also been a consistent source of support since my undergraduate days. While I do not get to see them nearly enough, it is always a joy to know that we can always pick up exactly where we left off.

Of course, I am incredibly indebted to mom, dad, grandma, John, Wendy, Jimmy, Wanyi, Frank, Chris, Danny, and Timmy, for their boundless love and unwavering encouragement.
## Contents

1 Introduction .................................................. 15
   1.1 Optimally Selecting Health Care Providers to Decrease Costs ... 16
     1.1.1 Problem of Interest ........................................ 16
     1.1.2 Our Contributions ........................................... 16
   1.2 Predicting Wait Time for Kidneys to Assist with Acceptance/Rejection Decisions for Waitlist Patients ......................... 17
     1.2.1 Problem of Interest ........................................... 17
     1.2.2 Our Contributions ........................................... 17
   1.3 Ranking Liver Patients by Disease Severity to Decrease Adverse Events on the Waitlist ........................................... 18
     1.3.1 Problem of Interest ........................................... 18
     1.3.2 Our Contributions ........................................... 18

2 Optimal Selection of Health Care Providers .......................... 21
   2.1 Introduction .................................................. 21
     2.1.1 Our Contributions ........................................... 23
     2.1.2 Outline ...................................................... 24
   2.2 The Dataset ................................................... 24
     2.2.1 Descriptive Analytics ........................................ 25
     2.2.2 Problem of Interest ........................................... 28
     2.2.3 Data Cleaning Procedures ..................................... 28
   2.3 Model Formulation ............................................. 30
     2.3.1 Creating Service Regions ..................................... 31
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.3.2 Notation</td>
<td>33</td>
</tr>
<tr>
<td>2.3.3 Mixed Integer Linear Optimization Approach</td>
<td>35</td>
</tr>
<tr>
<td>2.4 Insights from Computations</td>
<td>37</td>
</tr>
<tr>
<td>2.4.1 Computations Conducted</td>
<td>37</td>
</tr>
<tr>
<td>2.4.2 Computational Results and Sensitivity Analysis</td>
<td>38</td>
</tr>
<tr>
<td>2.4.3 Understanding Effects on Individual Procedures</td>
<td>45</td>
</tr>
<tr>
<td>2.5 Policy Implications</td>
<td>50</td>
</tr>
<tr>
<td>2.6 Future Directions and Conclusion</td>
<td>51</td>
</tr>
<tr>
<td>3 Accept or Decline? An Analytics-Based Decision Tool for Kidney Offer Evaluation</td>
<td>53</td>
</tr>
<tr>
<td>3.1 Introduction</td>
<td>53</td>
</tr>
<tr>
<td>3.2 Data and Methods</td>
<td>54</td>
</tr>
<tr>
<td>3.2.1 Data</td>
<td>55</td>
</tr>
<tr>
<td>3.2.2 Methods for Wait time Prediction</td>
<td>55</td>
</tr>
<tr>
<td>3.2.3 Observations and Independent Variables</td>
<td>55</td>
</tr>
<tr>
<td>3.2.4 Performance Evaluation</td>
<td>58</td>
</tr>
<tr>
<td>3.3 Results</td>
<td>60</td>
</tr>
<tr>
<td>3.3.1 AUC</td>
<td>60</td>
</tr>
<tr>
<td>3.3.2 Accuracy Within Clusters of Similar Patients</td>
<td>61</td>
</tr>
<tr>
<td>3.3.3 Example of Intended Usage</td>
<td>64</td>
</tr>
<tr>
<td>3.3.4 On the Adaptability of the Prediction Method</td>
<td>66</td>
</tr>
<tr>
<td>3.4 Discussion</td>
<td>67</td>
</tr>
<tr>
<td>4 Who Deserves Priority? Predicting Adverse Events on the Liver Transplant Waitlist</td>
<td>71</td>
</tr>
<tr>
<td>4.1 Introduction</td>
<td>71</td>
</tr>
<tr>
<td>4.2 History and Current Practice: Model for End-Stage Liver Disease (MELD)</td>
<td>72</td>
</tr>
<tr>
<td>4.2.1 The Pre-MELD Era</td>
<td>73</td>
</tr>
<tr>
<td>4.2.2 MELD</td>
<td>73</td>
</tr>
</tbody>
</table>
### List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-1</td>
<td>Total costs paid out by insurance plan for procedures and drugs over time</td>
<td>26</td>
</tr>
<tr>
<td>2-2</td>
<td>Distribution of costs over patients for each of the four plan years</td>
<td>27</td>
</tr>
<tr>
<td>2-3</td>
<td>Small counter-example against greedy consideration of neighbors</td>
<td>32</td>
</tr>
<tr>
<td>2-4</td>
<td>Cost changes under provider optimization</td>
<td>39</td>
</tr>
<tr>
<td>2-5</td>
<td>Variation in total cost under different values of $\delta$</td>
<td>40</td>
</tr>
<tr>
<td>2-6</td>
<td>Variation in total cost under different values of $\bar{\phi}$</td>
<td>41</td>
</tr>
<tr>
<td>2-7</td>
<td>Variation in total cost under different values of $\lambda$</td>
<td>43</td>
</tr>
<tr>
<td>2-8</td>
<td>Variation in total cost under different values of $D^k$</td>
<td>44</td>
</tr>
<tr>
<td>2-9</td>
<td>Histogram of Colonoscopy/Biopsy pricing in original dataset and in the</td>
<td>46</td>
</tr>
<tr>
<td></td>
<td>solution to the provider optimization problem</td>
<td></td>
</tr>
<tr>
<td>2-10</td>
<td>Histogram of distances traveled between original service location and new</td>
<td>47</td>
</tr>
<tr>
<td></td>
<td>provider prescribed by the optimal solution</td>
<td></td>
</tr>
<tr>
<td>2-11</td>
<td>Histogram of capacity utilization levels at providers delivering Colonoscopy/</td>
<td>49</td>
</tr>
<tr>
<td></td>
<td>Biopsy by multiplicative factor of original service level</td>
<td></td>
</tr>
<tr>
<td>2-12</td>
<td>Histogram of capacity utilization levels at providers delivering Colonoscopy/</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>Biopsy by absolute change in number of procedures delivered</td>
<td></td>
</tr>
<tr>
<td>3-1</td>
<td>Visualization of different training/validation/testing splits</td>
<td>60</td>
</tr>
<tr>
<td>3-2</td>
<td>Variable importance plot for three-month, KDPI $\leq 0.4$ specification</td>
<td>61</td>
</tr>
<tr>
<td>3-3</td>
<td>Variable importance plot for six-month, KDPI $\leq 0.4$ specification</td>
<td>62</td>
</tr>
<tr>
<td>3-4</td>
<td>Variable importance plot for twelve-month, KDPI $\leq 0.4$ specification</td>
<td>62</td>
</tr>
</tbody>
</table>
3-5 Predicted versus actual probabilities of a KDPI ≤ 0.4 offer within six months for Type O in California (CAOP) ........................................ 63
3-6 Predicted versus actual probabilities of a KDPI ≤ 0.4 offer within six months in California (CAOP) ........................................ 64
3-7 Predicted versus actual probabilities of a KDPI ≤ 0.4 offer for patients of blood group O in New York (NYRT) within six months, by DR mismatches ......................................................... 65
3-8 AUC value evolution over increasing amounts of data .................. 66
4-1 Variable importance plot for the 3-month adverse event prediction random forest model ......................................................... 83
4-2 Geography of service regions .................................................. 84
4-3 Mean number of adverse events averted by region in simulation by random forest, unadjusted MELD, and HCC-adjusted MELD for organ arrival sample paths from previous years ......................... 86
4-4 Percentage increase in adverse events averted by random forest and unadjusted MELD over HCC-adjusted MELD for all patients in each region for organ arrival sample paths from previous years ............. 87
4-5 Percentage change in adverse events averted by random forest and unadjusted MELD over HCC-adjusted MELD for cancer patients in each region for organ arrival sample paths from previous years ........ 87
A-1 Histogram of price changes for Upper GI Endoscopy Biopsy ........ 94
A-2 Histogram of travel distances for Upper GI Endoscopy Biopsy ........ 94
A-3 Histogram of price changes for CT Scan of Abdomen and Pelvis ... 95
A-4 Histogram of travel distances for CT Scan of Abdomen and Pelvis ... 95
A-5 Histogram of price changes for MRI of Brain ............................ 96
A-6 Histogram of travel distances for MRI of Brain ............................ 96
A-7 Histogram of price changes for Lesion Removal Colonoscopy ........ 97
A-8 Histogram of travel distances for Lesion Removal Colonoscopy ........ 97
List of Tables

2.1 Quintiles of patient costs by plan year in dollars 26

3.1 Out-of-sample AUC values for the random forest model 61

3.2 Predicted versus actual probabilities that patient J.S. will be offered a kidney of KDPI \leq 0.4 within the next six months 65

3.3 Out-of-sample AUC values for random forest and classification tree 70

3.4 Out-of-sample AUC values for random forest and logistic regression 70

3.5 Predicted versus actual probabilities that patient J.S. will be offered a kidney of KDPI \leq 0.4 within the next six months for random forest and logistic regression 70

4.1 Mean number of adverse events averted for all patients in simulation by random forest, unadjusted MELD, and HCC-adjusted MELD for organ arrival sample paths from previous years 84

4.2 Mean number of adverse events averted for cancer patients in simulation by random forest, unadjusted MELD, and HCC-adjusted MELD for organ arrival sample paths from previous years 85
Chapter 1

Introduction

This thesis addresses three pressing problems facing businesses, health care practitioners, and transplant patients in our nation today: the ballooning costs of providing health insurance for employees; the difficult decision of whether to accept or reject a current kidney offer in anticipation of a future higher-quality offer; and, the critical—and often life-or-death—decision of determining which patients should receive greater priority for liver transplantation. For each of these problems, the unifying element is the centrality and importance of data and data analytics as part of our approach. In the domains of insurance claims and organ transplantation, high-quality data has been collected for many years now, and, in each chapter of this thesis, we propose novel frameworks for how to harness the power of these existing datasets to arrive at actionable decisions that can add a significant edge over current practice. Our contributions have important implications in lowering health care costs for employers and improving patient outcomes in transplantation.

We provide a brief outline of each of these three problems, as well as our contributions, before expositing them in more detail in the subsequent chapters.
1.1 Optimally Selecting Health Care Providers to Decrease Costs

1.1.1 Problem of Interest

Health care expenditures in the United States (US) have been increasing at unsustainable rates for more than thirty years with no signs of abating. A significant proportion of increasing expenditures can be attributed to a lack of price transparency. We aspire to address the problem of rising health care costs for self-insured employers with access to employees’ claims data by identifying lower cost providers in an optimal way.

1.1.2 Our Contributions

We propose an approach based on mixed integer linear optimization that utilizes the data to generate decisions on how to optimally select providers to minimize cost, while respecting convenience for employees. We demonstrate that our approach is tractable for large-scale datasets and present the computational results on a real health care claims dataset.

Our data-driven, optimization-based approach leads conservatively to a 10% reduction of health care costs in a real-world study involving a large private US employer with 14,000 dependents. This is achieved by reassigning patients to different providers for a small number of procedures with a relatively small degree of employee inconvenience (an average (maximum) increase of health care related travel of 18km (40km)). These cost reductions are robust to changes in a variety of parameters.

Our framework generates a list of low-cost providers capable of providing an assortment of procedures that adequately cover the existing patient population. This list could be used to suggest providers or inform potential partnerships for service contracts. Our data-driven framework for reducing costs underscores the potential of harnessing the power of large datasets and associated approaches towards alleviating one of the most serious issues facing the US today.
1.2 Predicting Wait Time for Kidneys to Assist with Acceptance/Rejection Decisions for Waitlist Patients

1.2.1 Problem of Interest

When a deceased-donor kidney is offered to a candidate on the kidney transplantation waitlist, the decision to accept or decline the organ relies primarily upon a health care practitioner’s experience and intuition. Such decisions must achieve a delicate balance between estimating the immediate benefit of transplantation and the potential for future higher-quality offers. However, the current experience-based paradigm lacks scientific rigor and is subject to the inaccuracies that plague anecdotal decision-making. We aim to progress beyond this intuition-based decision-making process by providing data-driven support for practitioners who are responsible for these complex decisions. In this way, we aspire to significantly improve patient outcomes in kidney transplantation.

1.2.2 Our Contributions

We develop a data-driven analytics-based model to predict whether a patient will receive an offer for a deceased-donor kidney of varying quality thresholds at time frames of three, six, and twelve months. Our model accounts for the patient’s service region, blood group, wait time, antigen profile, and prior offer history to provide accurate and personalized predictions.

Using historical data from 2007 to 2013, we train our model and test it out of sample. Our model achieves an out-of-sample AUC of approximately 0.87 for all of the kidney quality thresholds and time frames that we consider for the ten most populous service regions. We investigate the adaptability of our model given increasing amounts of data and observe that as additional data becomes available, AUC values increase before subsequently leveling off.
Our data-driven analytics-based model may assist transplant practitioners and candidates during the complex decision of whether to accept or forgo a current kidney offer in anticipation of a future high-quality offer. In summary, our approach holds promise as a means to facilitate timely transplantation and optimize the efficiency of the allocation overall.

1.3 Ranking Liver Patients by Disease Severity to Decrease Adverse Events on the Waitlist

1.3.1 Problem of Interest

When a deceased-donor liver is procured, health care practitioners must make a determination as to which patients deserve the greatest priority for transplantation. Unlike kidney transplant candidates for whom the option for dialysis significantly reduces the risk of demise while waiting for the “next” offer, liver transplant candidates face a more immediate “life-or-death” situation whereby disease severity cannot be reliably controlled in the absence of a liver transplant. Current practice relies on ranking patients by their model for end-stage liver disease (MELD) score, which has been haphazardly tweaked and adjusted to compensate for deficiencies in overall waitlist mortality rate and fairness for particular patient subpopulations that have been observed \textit{ex-post}. We aspire to introduce a unified model that accounts for all patient characteristics to decrease adverse events on the waitlist and enable a more equitable organ allocation policy.

1.3.2 Our Contributions

Given historical data on the liver transplantation waitlist dating back almost two decades, we train a predictive model for the problem of determining the probability with which patients will suffer an adverse event of either death or unsuitability for liver transplant within the next three months. Our model not only greatly outperforms the MELD score and its variants in out-of-sample AUC for this particular prediction...
problem, but it also addresses the shortcomings of MELD and its variants when making accurate predictions for cancer patients.

We conduct a simulation experiment comparing the number of adverse events averted when ranking waitlist patients by the currently implemented MELD policy and our model, and we observe that our model increases the number of adverse events averted over the current MELD score by approximately 25% over the entire patient population while also maintaining fairness for cancer patients.
Chapter 2

Optimal Selection of Health Care Providers

2.1 Introduction

For more than thirty years, the cost of health care in the United States (US) has been increasing at a rate far greater than in other industrialized countries around the world [Squires, 2012]. In 1980, total health care expenditures accounted for approximately 9% of the total gross domestic product (GDP) of the US. By 2009, it accounted for 17.4% of total GDP. In contrast, the country with the second-highest health spending as a percentage of GDP in 2009 was the Netherlands, at 12% of total GDP. The ballooning of health care costs in the US is likely to continue its upward trajectory; unless action is taken, spending on health care is projected to reach a quarter of total US GDP by 2037 [Squires, 2012].

Per capita spending on health care (after adjusting for purchasing power parity) in the US was almost 80% higher than in Norway, the next highest country [Squires, 2012]. The greater expense of health care costs in the US has been attributed to higher prices than in other OECD countries. However, these higher prices have not translated to better quality. For instance, US hospitals are about average when measured by in-hospital mortality rates for heart attacks and stroke, and far below average for largely preventable deaths due to complications from asthma.
and diabetic amputations [Squires, 2012]. Additionally, the five-year survival rate for breast cancer, cervical cancer, and colorectal cancer is no higher in the US than in other OECD countries [Squires, 2012].

One of the main drivers of higher prices in the US health care industry is the lack of transparency of how much a procedure costs [Beck, 2014]. The amount paid for a particular procedure is highly dependent on who is paying the bill and can differ, in some instances, by several orders of magnitude. For example, Medicare would reimburse $335 for an MRI at a hospital in Dearborn, Michigan, while some private insurance companies would pay up to $1,990 for the same procedure at the same hospital [Beck, 2014]. Some of these large pricing differentials are a result of the asymmetry of market power between private health insurance companies and larger regional hospitals. Many experts also argue that prices vary as a result of different accounting methods that allow hospitals to recoup losses from more costly services, although this does not explain high costs for office visits [Beck, 2014].

In addition to the opaque nature of price setting for medical procedures, many large-scale providers have implemented non-competitive contracting policies that lead to unjustified price differentials [Commonwealth, 2011]. These providers, often spread out across large geographical areas, force payers to negotiate on a single price for all locations, even if payers only wish to utilize one location. This leads to higher prices in locations where geographic differences do not warrant them and exacerbates the problem of large variations in price within a small area.

Many experts in the medical community have started calling for full transparency of pricing for medical procedures [Emanuel et al., 2012, Sinaiko and Rosenthal, 2011]. While some states such as California, Massachusetts, and New Jersey have started reporting basic summary statistics regarding charges for common procedures in hospitals in an effort to empower patients as they decide where to seek treatment, some have argued this transparency in and of itself does not lead to decreased costs [Sinaiko and Rosenthal, 2011]. In fact, since most patients are insured and thus immunized against much of the price variation above their deductibles, it is possible for such transparency to lead to increased costs on the part of the insurer, as the
incentives for seeking cheaper care are severely reduced.

Many insurance companies have begun to develop reference prices for particular procedures; once reference prices are set, they will only cover costs up to the reference price. Patients are free to select their medical provider, but are responsible for any charges incurred above the reference price. The potential implications of data-driven pricing cannot be understated. Using reference pricing on hip and knee replacements alone, WellPoint, a large insurance company, was able to decrease prices in more expensive hospitals by up to one-third between 2008 and 2012 [Reinhardt, 2013]. With the advent of more plentiful electronic data, the time is ripe for employers to utilize data-driven decision-making to tackle the unsustainable explosion in health care costs in the US.

2.1.1 Our Contributions

In this paper, we propose a framework through which employers who self-insure can identify lower-cost medical providers that are convenient for their employees. Our framework utilizes historical electronic claims data to determine providers that are already charging reasonable prices and that are in close geographical proximity to the patient population. We discuss how to process general electronic claims datasets to extract the relevant information necessary to conduct this analysis.

Our main contributions are two-fold. First, we propose an approach based on mixed integer linear optimization that can be practically solved to optimality for large-scale datasets. We then demonstrate that the computational results returned by this approach can yield significant reductions in overall costs for medical procedures. We consider the sensitivity of our results to changes in a variety of our initial assumptions and observe that the cost savings are robust to these changes. Our modeling framework naturally lends itself to providing actionable decisions, thus bridging the gap between descriptive and prescriptive analytics. We also provide a discussion of how these prescriptions for optimally selecting medical providers can be realized in practice, so that these cost savings can be achieved.
2.1.2 Outline

In the next section, we discuss our dataset in more detail and provide some descriptive data analytics. We also formally define our problem of interest and detail our data cleaning procedures. In Section 3, we present a tractable mixed integer linear optimization approach that solves our problem of interest. In Section 4, we discuss our computational results and analyze the sensitivity of the results to different parameter settings. We also consider how the changes prescribed by our approach affect the prices paid for medical procedures, the number of procedures delivered by medical providers, and the distances that patients must travel. Section 5 delves into the policy implications of our work and addresses ways to implement changes in health insurance to achieve cost reductions for medical procedures. Section 6 concludes.

2.2 The Dataset

We obtained a dataset consisting of all health care claims paid out between 1 January 2011 and 1 July 2015 by a large self-insuring private employer based in the southern United States. Our dataset is 1.13GB in size, with 108 columns and approximately 1.55 million rows. Each row contains information about an insurance claim that was filed for a procedure delivered to an employee or dependent, which is described using the variables encoded in the columns. For the purposes of our analysis, we restrict attention to the following columns:

- **ClaimType**, a factor variable that denotes either medical procedure or prescription;
- **LocationType**, the type of facility where the procedure was provided (e.g., office, inpatient hospital);
- **PersonId**, a unique ID number for each individual patient;
- **DOSStart**, the start date of the medical procedure;
- **DOSEnd**, the end date of the medical procedure;
• **ICD9, Name, Category**, descriptions of the International Classification of Diseases (ICD9) code for the diagnosis;


• **ProcedureName**, the name of the medical procedure;

• **ProviderName**, the name of the provider who delivered the procedure;

• **ProviderNPI**, the provider’s unique National Provider Identifier (NPI);

• **AmtPaid**, the total amount paid for the procedure.

### 2.2.1 Descriptive Analytics

The dataset contains 1.55 million claims from more than 14,000 patients nationwide. These claims amounted to a total expenditure of more than $151 million, of which approximately $121 million were for medical procedures, and $30 million were for prescription drugs. Approximately 30,000 providers and 4,792 medical procedures appear in the dataset. Figure 2-1 breaks down the costs of procedures and drugs by plan year. Note that costs from plan year 2015 are lower than preceding years because our dataset only includes claims from the first half of the year. Furthermore, note that the plan year of a claim may not correspond to the actual year in which the procedure occurred or prescription was filled, due to delays in billing.

The mean total cost (for procedures and prescription drugs) per patient in plan year 2012 was $4032, increasing approximately 10% to $4473 in plan year 2013. Although there was a slight decrease in mean cost per patient in plan year 2014, halfway through plan year 2015, the mean total cost is $2415 per patient. This suggests that once all claims for 2015 are processed, there will be an increase in mean cost per patient by about 10% over that of plan year 2014. On the other hand, median cost per patient in plan year 2012 was $1011, with slight year-to-year decreases. Median cost per patient in plan year 2014 was $915. These summary statistics suggest that the distribution of patient costs is highly skewed.
Figure 2-1: Total costs paid out by the insurance plan for procedures and drugs over time. Note that our dataset ends in June 2015, so the visualization only captures half of Plan Year 2015.

We can take a deeper look into how costs per patient are distributed over the patient population by looking at the quintiles of the amount paid per patient over each plan year in Table 2.1.

<table>
<thead>
<tr>
<th>Percentile of Patient</th>
<th>Plan Year</th>
<th>20</th>
<th>40</th>
<th>60</th>
<th>80</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2012</td>
<td>$4,000</td>
<td>10,000</td>
<td>23,000</td>
<td>59,000</td>
<td>466,000</td>
</tr>
<tr>
<td></td>
<td>2013</td>
<td>5,000</td>
<td>13,000</td>
<td>30,000</td>
<td>88,000</td>
<td>521,000</td>
</tr>
<tr>
<td></td>
<td>2014</td>
<td>5,000</td>
<td>13,000</td>
<td>33,000</td>
<td>85,000</td>
<td>567,000</td>
</tr>
<tr>
<td>(*)</td>
<td>2015</td>
<td>3,000</td>
<td>8,000</td>
<td>20,000</td>
<td>45,000</td>
<td>283,000</td>
</tr>
</tbody>
</table>

Table 2.1: Quintiles of patient costs by plan year in dollars. Most patients incur relatively low cost, while a small percentage of the patients account for a large portion of total expenditures. (*) Data available only for the first half of 2015.

Figure 2-2 shows the cumulative contribution to the total amount paid in each plan year after sorting all of the patients by their costs. For instance, we can see that in each of the plan years, the lowest 50% of all patients contribute only about 5% of the overall cost for that particular year. The top 10% of the most expensive patients contribute approximately 60% of all of the costs. The black line in the figure is a reference for perfectly equal distribution of costs over all of the patients. Overall, from this figure, we can see that the distribution of costs over the patients is highly
unequal in any particular plan year and that this distribution remains roughly the same from year to year.

Figure 2-2: Distribution of costs over patients for each of the four plan years. The black diagonal line represents perfect equality among all patients. The distribution of costs per patient is highly skewed, and we can observe that the top 5% of patients by amount paid account for a majority of healthcare expenditures. We also note that this distribution stays roughly the same from year to year.

Both Table 2.1 and Figure 2-2 corroborate findings in other studies [Stanton, 2005]. In particular, we see that the top 5% of patients by amount paid accounted for approximately 55% of all costs in any given year. Conversely, the bottom 50% of patients by amount paid accounted for about 4% of all costs. This suggests that our dataset is reflective of the US population at large and indicates that our results could reasonably be extrapolated to a larger scale.
2.2.2 Problem of Interest

We aspire to identify a set of low-cost medical providers who can deliver a wide range of procedures within close geographical proximity of the patient population. Because our dataset includes information on how much was paid to medical providers for particular procedures in the past, we can identify providers who have costs that are consistently lower than other providers delivering the same procedure. In this way, we allow the data to inform our decisions as to which providers are already charging reasonable prices for the same procedures. Our data also gives us a good sense of the historical demand that these providers have satisfied from our patient population.

Given this data, our goal is to understand the magnitude of potential savings and to identify providers who could help realize these reductions in cost. This collection of providers and their prices for procedures could serve as the starting point for further negotiations leading to even lower costs. We present the particular data cleaning procedures that are necessary to extract the inputs specific to our optimization model in the next section.

2.2.3 Data Cleaning Procedures

The main aspiration of our work is to identify lower cost providers who are capable of delivering the same procedures while remaining in close geographical proximity to the patients’ original providers. As a result, the two goals of our data cleaning process are to identify the actual locations where procedures occurred as well as the cost distribution for a specific provider-procedure pair.

Identifying Procedure Delivery Locations: Because we wish to understand price discrepancies for the same procedure within a particular geographic vicinity, our analysis relies on the accuracy of where particular procedures were provided.

An issue specific to claims datasets is that locations within the dataset often refer to billing addresses, as opposed to the actual location where medical services were delivered. Using an additional dataset provided by the Centers for Medicare and Medicaid Services that specified the locations of medical providers based on
their National Provider Identifier (NPI), we were able to obtain the actual provider location [Medicare, 2015]). Although this allowed us to properly identify the zip codes of practice locations, many of the rows in our dataset did not have valid NPI information. There was no additional information in the dataset from which we could obtain the practice locations of these claims, so we discarded them for the purposes of our analysis. Of the original 1.5 million rows, we discarded 100,000 rows (approximately 6%). These removed rows accounted for approximately 10% of the cost in the entire dataset.

One additional challenge is that a single procedure might generate multiple claims with different delivery locations. Upon further examination, we noticed that for almost all of these instances, one of the claims was billed by a physician or specialist, while the other was billed by a hospital or medical center, indicating that a medical provider was utilizing nearby hospital facilities. To obtain the correct delivery location, we concluded that the actual delivery zip code of the medical procedure was in the hospital or medical center. We were able to perform this location finding procedure by utilizing the LocationType column in the dataset and specifying a precedence ordering over the location types (e.g., outpatient hospital takes precedence over office).

It is important to note that this data cleaning step is also crucial in determining the true cost distribution for a specific provider-procedure pair, which we describe below.

Obtaining Costs for Provider-Procedure Pairs: Without properly understanding the true cost distribution for a specific provider-procedure pair, our approach might select providers based on prices that are too low, thus affecting the validity of our results.

Two challenges involved in obtaining an accurate distribution of costs for all provider-procedure pairs are procedures involving multiple providers and refunded claims. As we discovered while identifying procedure delivery locations, a single procedure provided to a single patient can often result in multiple claims. Thus, it is not appropriate to aggregate rows simply by ProviderName and ProcedureName, because
multiple providers may have taken part in the same procedure. We instead identify all of the instances in which a single procedure led to multiple claims and we aggregate these rows individually. When we aggregate these rows, we replace them with a single row containing the sum of all of the costs paid for that procedure. We also create a new provider that is a combination of all of the providers who took part in the procedure. As before, we use location type precedence ordering to determine the true delivery location of the procedure.

Both of these data cleaning steps present several unique challenges and are crucial to the validity of our results. Overall, our data cleaning procedures reduced the number of rows in our dataset from 1.5 million to about 520,000 and accounted for approximately $82.5M in costs.

2.3 Model Formulation

After completing the data preparation procedures from the previous section, each row in our dataset is structured as follows:

\[
\text{claim} = (\text{PersonId}, \text{ProcedureName}, \text{ProviderName(s)}, \text{LocationZIP}, \$\text{Paid})
\]

We think about our problem as follows. If a patient requires procedure \( k \) in a particular zip code \( z \), which provider is within close proximity of \( z \) and can provide \( k \) at a low cost? For our analysis, we formulate a model whose solution represents perfect hindsight; that is, given the historical claims data that we have collected so far, we consider the optimal re-assignment of patients to nearby low-cost providers. As we discuss in the next section, most of the claims in the dataset are concentrated within a small geographical region. This allows us to disaggregate the optimization problem into smaller, separable subproblems without sacrificing optimality.
2.3.1 Creating Service Regions

We restrict our attention to claims related to medical procedures. Additionally, we define a maximum acceptable travel distance that denotes a radius of how far we will are willing to look for lower-cost providers. For the purposes of our analysis, we have selected a maximum acceptable travel distance, which we denote by $\delta$. We conduct sensitivity analysis on an appropriate selection of $\delta$ between 30 and 50 kilometers, a distance that patients can reasonably travel within 20-45 minutes.

Because we do not look for providers beyond a distance $\delta$ from the original service location, we can divide our problem into separable regions, where each region consists of zip codes that occur within our dataset. If we were to draw an edge between every pair of zip codes within $\delta$ from each other, a region would be equivalent to a connected component in the graph. To create our regions, we successively grow out each region until no additional zip codes within $\delta$ can be reached. This procedure produces a set of disjoint regions $r \in R$.

Given these regions, we can decompose the overall problem by solving a single optimization problem for each $r \in R$. By construction, for any $z, z' \in Z$ such that $z \in r, z' \in r', r \neq r'$, observe that any changes to the capacity or price of the provider at $z'$ do not change our decision with respect to optimizing over region $r$. Thus, when identifying potential low-cost providers for an existing service location, we can solve our optimization problem separately over each of the individual regions, without sacrificing optimality.

In contrast, it is not appropriate to simply consider a zip code $z$ and its neighboring regions $N(z)$ in isolation via a greedy approach, because for a particular zip code $z$, changes in capacity or pricing at some zip code that is two hops away may change our decision when optimizing over $r \ni z$. We present a simple counterexample in Figure 2-3.

In the counterexample provided in Figure 2-3, we have three zip codes where the same procedure can be obtained. Zip code A is within $\delta$ distance of B and C, but the distance between B and C is greater than $\delta$. Nonetheless, the price of the procedure
optimization problem separately over each of the individual regions, without sacrificing optimality. In contrast, it is not appropriate to simply consider a zip code \( z \) and its neighboring regions \( N(z) \) in a greedy approach, because for a particular zip code \( z \), changes in capacity or pricing at some zip code that is two hops away may change our decision when optimizing over \( r_z \). We present a simple counterexample in Figure 4.

![Figure 4: A small counter-example against greedy consideration of neighbors. Provider A is within \( \delta \) of Providers B and C; Providers B and C are more than \( \delta \) apart. However, a change in the price of Provider C from 8 to 4 will lead to changes in the utilization of Provider A.](attachment:image.png)

at zip code C determines how much of the capacity at B will be used in the optimal solution. For instance, at the current pricing levels, the provider in zip code B will serve all patients in zip codes A and B, while the provider in zip code B serves all patients in zip code C. However, if the price for a procedure at zip code C is decreased to 4, then the provider in zip code B serves only patients from that zip code, while the provider at zip code C serves all patients from zip codes A and C. Thus, we see that even though zip code C is not within \( \delta \) distance of zip code A, the data related to zip code C still affects the optimal solution in this subset of the graph.

We can think of our optimization approach in terms of a graph problem. Let \( J \) denote the set of providers, and \( K \) denote the set of all procedures. Each provider is a node with edges representing providers in nearby regions. Each node, indexed by \( j \in J \), has some level of demand for procedure \( k \in K \), denoted \( D^k_j \). Correspondingly, each node has some maximum capacity \( C^k_j \). A pair of nodes \( j, j' \) are connected by an edge if the zip codes of their service locations are within \( \delta \) of each other. Furthermore, suppose each provider \( j \) has a procedure price of \( \mu^k_j \). Our problem becomes related to a minimum vertex cover problem where we must select a collection of nodes (providers) that have the capacity to cover the demand of all neighboring unselected nodes at minimum price.
For this particular application, we focus our attention on the largest region created using $\delta = 50\text{km}$. Since the workforce for this particular employer is largely centralized in a metropolitan area in the southern United States, it is not surprising that approximately $77\%$ of the costs for medical procedures in this dataset fall within this largest region. Of the almost $520,000$ rows related to medical procedures in the cleaned dataset, more than $400,000$ occurred within this largest region. The total expenditure over the $3.5$ years observed in the dataset was approximately $63.6\text{M}$. In all, there were more than $21,000$ providers belonging to this region in our dataset. Additionally, more than $3,600$ different types of procedures were billed in this region.

### 2.3.2 Notation

Before we present our optimization approach, we introduce the notation that we use for our data, parameters, and variables.

**Data:**

- $k \in K$, the set of procedures;
- $j \in J$, the set of providers, each belonging to a single zip code $z$;
- $z \in r$, the set of zip codes in the region of interest $r$.
- $N(z) = \{z' | \|\text{loc}(z) - \text{loc}(z')\| \leq \delta, z' \in r\}$, the set of zip codes (including $z$ itself) that are neighbors of $z$;
- $D^k_j$, the existing demand for procedure $k$ at provider $j$;
- $\mu^k_j$, the average cost for procedure $k$ at provider $j$;
- $C^k_j$, the capacity for procedure $k$ at provider $j$.

Since we are re-assigning patients in hindsight, we obtain demand information $D^k_j$ based on the actual number of times provider $j$ delivered procedure $k$ in our claims dataset. We obtain the average price charged by provider $j$ for procedure $k$ in the same way. To protect ourselves against data error, we set a lower limit for the average
cost $\mu_{ij}^k$ based on Medicare reference pricing; if the average price for procedure $k$ at provider $j$ is lower than the Medicare price, we conservatively estimate that we can obtain the Medicare reference price for that geographical region.

We do not have capacity data for individual providers, so we assume that each provider can deliver up to some multiplicative factor of the historical demand from our dataset. For example, if a hospital has delivered 10 knee replacements to our patient population in the last three years, we assume that they could have delivered up to 30 knee replacements. Thus, we create a new parameter $\overline{D}^k$ that represents the maximum allowable increase under demand reallocation for a particular procedure $k$. We represent the capacity $C_{ij}^k$ as a the product $\overline{D}^k D_j^k$ for varying levels of $\overline{D}^k$. We test the sensitivity of the optimal solution to varying levels of $\overline{D}^k$ in Section 4.

We also include additional constraints that limit the total number of different types of procedures that we are willing to reallocate. This ensures that the optimal solution does not go too far in terms of reallocating all of the possible procedures that could be provided. Instead, it allows us to focus on the procedures where the maximum cost savings can be achieved. The optimal solution provides this list of procedures and could form the basis for further negotiation of prices with medical providers. We test the sensitivity of the optimal solution to varying levels of this parameter, which we denote by $\overline{\phi}$.

**Parameters:**

- $\lambda$, the objective penalty for each additional provider that we contract with;
- $\overline{\phi}$, the maximum number of types of procedures that are reallocated;
- $\overline{D}^k$, the maximum allowable increase under demand reallocation for a particular procedure $k$;
- $\delta$, the acceptable distance between existing and new providers.

**Variables:**

- $n_{ij}^k$, the number of patients assigned to receive procedure $k$ at provider $j$;
• $\phi^k_{z,z'}$, the amount of demand for procedure $k$ transferred from zip code $z$ to $z'$;

• $\phi^k \in \{0, 1\}$, which takes value 1 if we allow demand for procedure $k$ to be reallocated, 0 otherwise;

• $x^k_j \in \{0, 1\}$, takes value 1 if we select provider $j$ to deliver procedure $k$, 0 otherwise;

• $y_j \in \{0, 1\}$, takes value 1 if we choose provider $j$ for at least one procedure, 0 otherwise.

### 2.3.3 Mixed Integer Linear Optimization Approach

Using the above notation, we formulate the following mixed integer linear optimization problem (Formulation 1):

\[
\begin{align*}
\min_{n,x,y,\theta} & \quad \sum_{j,k} n^k_j \mu^k_j + \lambda \sum_j y_j \quad \text{(minimize cost + # of providers)} \quad (2.1) \\
\text{subject to} & \quad \sum_{z' \in N(z)} \phi^k_{z,z'} \geq \sum_{j \in z} D^k_j, \quad \forall k, z, \quad \text{(demand outflow)} \quad (2.2) \\
& \quad \sum_{z' \in N(z)} \phi^k_{z,z'} \leq \sum_{j \in z} n^k_j, \quad \forall k, z, \quad \text{(demand inflow)} \quad (2.3) \\
& \quad \sum_k \phi^k \leq \bar{\phi}, \quad \text{(limit procedure transfer)} \quad (2.4) \\
& \quad n^k_j \leq D^k_j + M^1_{j,k} \phi^k, \quad \forall j, k, \quad \text{(enforce procedure limit)} \quad (2.5) \\
& \quad n^k_j \geq D^k_j - M^2_{j,k} \phi^k, \quad \forall j, k, \quad \text{(enforce procedure limit)} \quad (2.6) \\
& \quad n^k_j \leq \bar{D}^k_j D^k_j x^k_j, \quad \forall j, k, \quad \text{(respect capacity)} \quad (2.7) \\
& \quad y_j \geq x^k_j, \quad \forall j, k, \quad \text{(number of providers)} \quad (2.8) \\
& \quad x^k_j, y_j \in \{0, 1\}, \quad \forall j, k, \quad (2.9) \\
& \quad \phi^k_{z,z'} \geq 0, \quad \forall k, z, z'. \quad (2.10)
\end{align*}
\]
where \( M_{1,j,k} = D_j^k(D_k^k - 1) \) and \( M_{2,j,k} = D_j^k \) so that constraints (2.5, 2.6) are enforced.

The objective function (2.1) is to minimize the total costs incurred if we were to optimally re-assign patients to providers, with a penalty of \( \lambda \) for each provider that we select. This objective penalty is added so that the optimal solution does not select too many providers. From a convenience perspective, a small number of providers is desirable because it means that patients need not visit too many providers for their health care needs. Furthermore, it allows us to identify providers that have reasonable pricing for a wide range of procedures. It is notable that the solution will not necessarily select providers who have the lowest price for a particular procedure; it may select providers who have above average costs on some procedures, but low costs on others.

The constraints (2.2) and (2.3) can be thought of as flow constraints that match the original demand to the re-assigned providers via conservation of flow. Notice that the variable \( \theta_{k,z',z}^z \) keeps track of the amount of demand that is transferred from zip code \( z \) to neighboring zip code \( z' \). In constraint (2.2), this aggregates the outflow of demand from a particular zip code \( z \) to its neighboring zip codes. In constraint (2.3), we consider the inflow of demand into a particular zip code \( z \) from its neighboring zip codes. Since ultimately the variable \( n_j^k \) represents the amount of procedure \( k \) that is delivered by providers \( j \) that are located within \( z \), this constraint means that the demand flowing in must be covered.

Constraint (2.7) enforces capacity constraints for provider-procedure pairs, while constraint (2.8) allows the variable \( y_j \) to count the total number of providers that we select.

This formulation has the desirable property that it does not create too many constraints because it aggregates demand flow into the variable \( \theta \). In particular, constraints (2.2) and (2.3) need only be created for \( (k, z) \) pairs that exist in the data; that is, (provider, zip code) pairs. Since the (provider, zip code) matrix is very sparse, this greatly reduces the number of constraints that we need. We also note that even though \( \theta \) is indexed by \( k, z, z' \), each individual zip code \( z \), the size of the neighboring zip codes \( N(z) \) is very small, so we need not create too many variables. Overall,
even though the largest region contains more than 1,300 zip codes, more than 3,600 procedures, and over 21,000 providers, the formulation requires only approximately 177,000 constraints and 184,000 variables.

2.4 Insights from Computations

We implemented Formulation 1 with data from the largest region in our dataset in Julia and the mathematical programming package JuMP [Lubin and Dunning, 2015]. The model was solved using Gurobi 6.5.0 as the mixed integer linear optimization solver. We ran all computations on a desktop equipped with an Intel Core i7-3770 quad-core CPU running at 3.4GHz and 32GB of RAM.

2.4.1 Computations Conducted

We solved Formulation 1 for the following ranges of parameter values:

- \( \lambda \in \{100, 200, 500, 1000, 2000, 5000, 10000, 20000\} \), the penalty for each additional provider selected;
- \( \bar{\phi} \in \{25, 50, 100, 200, 500\} \), the maximal number of types of procedures that we are allowed to reallocate;
- \( \bar{D}^k \in \{2, 2.5, 3\} \), the maximal allowable increase in demand for procedure \( k \). Where existing demand for procedure \( k \) at provider \( j \) is \( D_j^k \), we allow the new maximal capacity to be equal to \( C_j^k = \bar{D}^k D_j^k \);
- \( \delta \in \{30, 40, 50\} \) km, the maximum allowable travel distance for the patient.

We set a time limit of 600 seconds for each run. Over each of the different parameter settings, we noticed that high quality solutions with optimality gap of less than 1% can be found in less than two minutes. In 240 of the parameter settings, provable optimality was achieved within the time limit. Of the remaining 120 cases, the largest optimality gap at the time limit of 500 seconds was 0.23%. Notably, all
of the cases that did not reach provable optimality had large values ($\geq 2000$) of $\lambda$. This is due to the fact that $\lambda$ implicitly defines a maximum number of providers that can be selected. Larger values of $\lambda$ lead to fewer providers, making the problem more difficult to solve because the optimal solution must trade off between selecting an additional provider with minimizing overall cost. In the limit where $\lambda = 0$, the optimal solution would simply select the providers that provide each individual procedure at the lowest cost in a greedy fashion.

### 2.4.2 Computational Results and Sensitivity Analysis

We present a visualization of the overall cost reduction under all of the different parameter values in Figure 2-4.

In Figure 2-4, we have converted the different values of $\lambda$ to the total number of providers that are selected in the optimal solution, which is presented on the horizontal axis. The vertical axis denotes the total cost over the three and a half years in the database under the optimal solution provided by Formulation 1. The black horizontal line denotes the original cost of approximately 63.6 million dollars that was incurred in the data. For additional reference, more than 21,000 providers were paid for delivering procedures in the original dataset.

Each point presented in Figure 2-4 denotes a particular set of values for the capacity increase $D^k$, the number of types of procedures reallocated $\phi$, a maximum travel distance $\delta$, and penalty $\lambda$ for each additional provider. The color of the point denotes the value of $D^k$, the level of transparency denotes the value of $\phi$, and the shape denotes the value of $\delta$. We see that, all else being equal, larger values of $\delta$, $D^k$, or $\phi$ lead to fewer providers selected as well as lower total cost. This is an intuitive result, because larger values of $\delta$ allow for increased flexibility in selecting replacement providers. As $D^k$ increases, we require fewer providers to deliver the same number of procedures, while also yielding additional savings from lower cost providers. Larger values of $\phi$, or allowing more types of procedures to be reallocated, also leads to fewer providers being selected, since the optimal solution will consolidate the delivery of procedures to a smaller set of providers. It also takes additional advantage of lower
Figure 2-4: Cost changes under provider optimization. Different maximum travel distances $\delta$ are denoted by different shapes; different values of capacity increase $D^k$ are denoted by the different colors; different values of $\phi$ are denoted by the different levels of transparency. All else being equal, larger values of $\delta$, $D^k$, or $\phi$ lead to both lower cost and fewer providers.

cost providers for a greater number of procedures.

From this visualization, we see that under some reasonable settings of parameter values, for instance, $\delta = 40$, $D^k = 2.5$, and $\phi = 100$, we can achieve a total cost of approximately 57 million dollars, which would represent a savings of more than 10%. At the same time, the number of providers selected to deliver procedures is
reduced by more than half. To further understand the effects of varying each of these parameters, we restrict our attention to the cases where three of the four parameters are fixed, and observe how varying the fourth parameter affects the total cost.

**Sensitivity to Maximum Allowable Travel Distance**

To better understand the effects of varying the maximum allowable travel distance $\delta$, we fix $\lambda = 1000$, $\overline{D}^k = 2$, $\phi = 100$, while varying $\delta \in \{30, 35, 40, 45, 50\}$ km. The total cost under each of these parameter settings can be seen in Figure 2-5.

![Cost Changes by Maximum Allowable Travel Distance](image)

Figure 2-5: Variation in total cost under different values of $\delta \in \{30, 35, 40, 45, 50\}$ km, with $\overline{D}^k = 2$, $\lambda = 1000$, and $\phi = 100$. We observe a somewhat linear decrease in total cost as the maximum travel distance increases.

We observe that cost decreases linearly as the maximum travel distance increases. This agrees with our intuition that allowing more flexibility in finding replacement...
providers will yield additional cost savings.

**Sensitivity to Number of Types of Procedures Reallocated.**

To better understand the effects of varying the number of different types of procedures reallocated, or the value of \( \overline{\phi} \), we fix \( \lambda = 1000 \) and \( D^k = 2 \) while varying \( \overline{\phi} \in \{10, 50, 100, 200, 500, 1000\} \). The total cost under each of these parameter settings can be seen in Figure 2-6.

![Cost Changes by # of Types of Procedures Reallocated](image)

Figure 2-6: Variation in total cost under different values of \( \overline{\phi} \in \{10, 50, 100, 200, 500, 1000\} \), with \( \delta = 40, D^k = 2, \lambda = 1000 \). We observe that cost reductions start slowing significantly after approximately 200 types of procedures being reallocated.

We observe that cost reductions start slowing significantly after reassigning approximately 200 different types of procedures out of a total of more than 3,600. This
comports with our intuition that there are a few medical procedures on which the majority of cost savings can be realized. Our approach extracts the names of the procedures that can yield the greatest cost savings via reallocation of patients to lower cost providers. Note that the particular choice of $D^k$ and $\lambda$ do not affect the overall shape of this curve too much, and that fixing them allows us to better isolate the effects of only varying $\phi$.

Increasing $\phi$ from 500 to 1,000 yields additional cost savings of less than half a million dollars, and further increasing $\phi$ past 1,000 leads to significantly diminished marginal savings. Since implementing these changes in provider selection are costly in practice, we would select approximately 100 to 200 different types of medical procedures on which to help patients select their providers. Even under the conservative setting of $D^k$ and $\lambda$, we observe an approximate cost savings of 10% for values of $\phi$ between 100 and 200.

**Sensitivity to Number of Providers Selected.**

To better understand the effects of varying the total number of providers selected to deliver medical procedures to our patient population, we vary the value of $\lambda \in \{100, 200, 500, 1000, 2000, 5000, 10000\}$, while fixing $\delta = 40, D^k = 2, \phi = 100$. Since $\lambda$ is a penalty in the objective of the optimization model for each additional provider selected, each value of $\lambda$ corresponds implicitly to a total number of providers. We present the cost changes under each of these different number of providers in Figure 2-7.

We observe that setting $\lambda = 1000$ yields a reasonable tradeoff between total number of providers selected and total cost savings. This corresponds to working with approximately 10,000 providers, which is a reduction in the original number of providers by more than half. Again, with conservative settings of $D^k = 2$ and $\bar{\phi}$, this value of $\lambda$ would bring about an almost 10% decrease in total costs related to medical expenditures in this service region.
Original Cost = $63.6M

Total Cost over 3.5 years (in millions)

Number of Providers

Cost Changes by Number of Providers Selected

Figure 2-7: Variation in total cost under different values of $\lambda \in \{100, 200, 500, 1000, 2000, 5000, 10000\}$, with $\delta = 40, D^k = 2, \phi = 100$. Note that each value of $\lambda$ implicitly determines the number of providers selected. We observe that cost reductions start slowing significantly after selecting about 10,000 providers. This corresponds to setting $\lambda = 1000$.

**Sensitivity to Capacity Assumptions.**

To better understand the effects of varying the assumed additional capacity that each provider could sustain, we vary $D^k \in \{2.0, 2.1, 2.2, \ldots, 3.0\}$ while fixing $\delta = 40, \phi = 100, \lambda = 1000$. We plot the results in Figure 2-8.

Here, we see a different change in total cost for each increment in factor of capacity increase. Notice that there are two significant drops in total cost at $D^k = 2.5$ and $D^k = 3.0$. This is due to the discrete nature of mixed integer optimization. It suggests that for the majority of providers that we are selecting, they only provide
Figure 2-8: Variation in total cost under different values of $D^k \in \{2.0, 2.1, 2.2, \ldots, 3.0\}$, with $\delta = 40, \phi = 100, \lambda = 1000$. Notice that there are significant jumps between 2.4 and 2.5, as well as between 2.9 and 3.0. This is due to the discrete nature of our optimization approach. Many of the lower-cost providers currently only deliver one or two procedures. Thus, these particular changes in $D^k$ lead to additional cost savings. For example, allowing a capacity increase of $D^k = 2.5$ for a provider currently delivering 2 procedures increases maximum capacity to 5. For $D^k = 3$, providers currently delivering 1 procedure have an increased maximum capacity of 3.

some procedures once or twice in our dataset. We explore this further in Section 2.4.3.
2.4.3 Understanding Effects on Individual Procedures

We now focus our attention on the results in the case where $\delta = 40$, $\overline{D}^k = 2.5$, $\bar{\phi} = 100$, and $\lambda = 1000$. Under these parameter values, the total cost incurred is 56.8 million dollars, or approximately 11% less than the original cost of 63.6 million dollars. The total number or providers selected is slightly less than 10,000 providers, representing a 50% decrease in the number of providers used. While this aggregate information is useful for understanding the overall effects of following the prescriptions of our approach, we also investigate the effects of reallocation on prices and traveling distances for patients of particular procedures to uncover additional insight into how these cost savings are achieved. We consider the additional burden on the selected providers by considering how many more procedures they would deliver under the reallocated regime. Our analysis shows that patients need not travel far from their original service locations to yield significant cost reductions. At the same time, most existing providers are only asked to deliver a few more procedures over their current service level.

Changes in Pricing

The optimal solution recovers the names of providers and procedures that can yield the most cost savings overall. One particular example on which we will focus is the procedure Colonoscopy/Biopsy. In this dataset, reallocating patients requiring this procedure can bring about a total of almost $129,000 in savings. In the original dataset, 284 patients received this procedure at an average cost of $1279. After optimally selecting providers, the average cost of this procedure is $825 per patient. Although we restrict our attention to Colonoscopy/Biopsy, it is important to note that similar observations can be made for many of the other procedures in the dataset. Figure 2-9 shows the distribution of the costs for Colonoscopy/Biopsy in the original dataset and in the optimal solution. We supply additional figures for other procedures in Appendix A.

From Figure 2-9, we can see that most Colonoscopy/Biopsy procedures are less
Figure 2-9: Histogram of prices for colonoscopies in the original dataset and in the solution to the provider optimization problem. We observe that almost all of the procedures in our optimized regime cost less than $1,200. In the historical dataset, approximately one-third of all patients received this procedure from providers charging more than this price.

than $1,200 in the optimized regime, while more than one-third of procedures exceeded this cost threshold in the original dataset. Under the optimal reallocation of patients to providers, we see that there is significantly less variation in prices for the procedure Colonoscopy/Biopsy. There is still some variation in price due to the geographical proximity and capacity increase constraints that we have placed on the problem. However, the optimal solution is effective at identifying costly providers and redirecting patients to providers of lower cost. It is impressive that under the reallocated regime, the 95th percentile of cost is comparable to the median price in the original dataset.
Traveling Distances

The procedure Colonoscopy/Biopsy is a relatively common procedure in the dataset, so our optimization approach is able to find many lower cost providers within 40 kilometers of the original service locations. Note that not all patients must travel the maximum allowable distance under optimized regime. In fact, as we can observe from Figure 2-10, a majority of patients need not leave the zip code of their existing provider under the optimized regime. Of the patients who must travel outside of their original zip code, the mean and median travel distance are about 18.1 and 13 kilometers, respectively.

![Histogram of Distances Traveled for Colonoscopy/Biopsy](image)

Figure 2-10: Histogram of distances traveled between original service location and new provider prescribed by the optimal solution. Of the 284 patients requiring Colonoscopy/Biopsy, 141 remained in the same zip code or did not need to travel. For patients who had to travel outside of the zip code of their original service location, the median travel distance is 13 kilometers, and the mean travel distance is 18.1 kilometers.

In practice, the parameter $\delta$ that denotes the maximum allowable travel distance can be varied. Larger values of $\delta$ would allow increased flexibility for provider reassignment, at the expense of decreased convenience to patients. To make our approach more general, we could add a penalty to the objective for each additional kilometer.
that a patient must travel, thus capturing this tradeoff more explicitly. In certain instances, such as very expensive procedures, it may be reasonable for patients to travel slightly farther, while for cheaper procedures, the cost savings do not warrant the additional inconvenience to the patient.

In some extreme cases, price discrepancies may be so large as to warrant treatment in an entirely different region. For instance, a group of 40 surgeons who believe strongly in price transparency opened the Surgery Center of Oklahoma, where all-inclusive prices for surgeries are posted openly on their website [Rosenberg, 2013]. These prices are so attractive to patients and their insurers that many patients travel across the country to receive treatment there. For the complicated procedure Lumbar Fusion, they list a price of $49,625 on their website. If the 28 patients in our dataset that received this particular procedure had sought treatment in Oklahoma, the total cost savings would be approximately $380,000. In this particular case, it may make sense for patients to incur some non-trivial cost of travel to attain more attractive pricing.

Changes in Capacity

We now consider the impact of the reallocated patient demand from the provider’s perspective. We have assumed that providers are capable of providing up to $D^k$ times their original service capacity. Under the optimal regime with parameters $\bar{D}^k = 2.5, \varphi = 100, \lambda = 1000$, we observe that our optimal solution utilizes all of the assumed extra capacity of 2.5 times the original service capacity at 10 out of the 149 providers that delivered Colonoscopy/Biopsy in the dataset. A histogram of utilization levels by medical provider is presented in Figure 2-11.

Figure 2-11 does not tell the whole story, however. This is because many of the providers in our existing dataset only appear once. As a result, simply adding one additional patient under the optimal regime would represent a 2.0x utilization factor over current service levels. This explains the jump in cost reduction for $\bar{D}^k = 2.5$ and 3.0 that we saw in Figure 2-8. Figure 2-12 presents the absolute change in the number of Colonoscopy/Biopsy delivered across all of the providers. We see
Figure 2-11: Histogram of capacity utilization levels at providers delivering Colonoscopy/Biopsy by multiplicative factor of original service level. Although 10 out of the 149 providers are assigned to deliver the maximum allowable number of procedures, 66 of the providers are no longer used. 29 providers remain at their current service levels.

that, with the exception of 7 providers, all existing providers are asked to deliver no more than 5 additional Colonoscopy/Biopsy procedures to reallocated patients. The seven remaining providers delivered 5, 6, 7, 8, 8, 10, and 10 Colonoscopy/Biopsy procedures in the original dataset, and are asked to perform 12, 15, 15, 17, 20, 25, and 25 procedures, respectively, under the optimal solution.

These findings are not specific to this particular procedure. We present similar figures for other procedures such as Upper GI Endoscopy Biopsy, CT of Abdomen and Pelvis, MRI of Brain, and Lesion Removal Colonoscopy in Appendix A.

For all of the procedures reallocated in the optimal solution, we did not notice an absolute increase of more than 50 patients per year for any given procedure at any individual medical provider. This suggests to us that the assumption of a 2.5x increase in capacity utilization over the existing demand levels from our dataset is reasonable. In addition, medical providers might welcome additional demand for procedures, and may even agree to preferred pricing if minimum volume guarantees
Figure 2-12: Histogram of capacity utilization levels at providers delivering Colonoscopy/Biopsy by absolute change in number of procedures delivered. All except for four of the providers are asked to deliver no more than 5 additional Colonoscopy/Biopsy procedures under the new allocation. The four providers in the right-most portion of the histogram are each asked to provide the maximum allowable increase in the optimal solution.

are achieved. In our view, our data-driven optimization approach could achieve cost reductions in practice by identifying specific procedures with potential for greatest savings. It also prescribes a set of low-cost medical providers within close proximity to patients’ existing service locations with whom an employer could contract for even greater price reductions.

2.5 Policy Implications

Our computational results suggest significant cost reductions can be achieved through relatively simple means. Due to the large pricing discrepancies between providers delivering the same procedures within small geographical neighborhoods, patients need not be inconvenienced too much under the reassignment prescribed by the optimal solution. Indeed, our results suggest that there is ample opportunity for cost reduction;
our results are just the first step in identifying all of the opportunities for decreasing an employer’s expenditures in providing health care for its employees.

One immediately actionable policy would be to compile a list of low-cost providers that are capable of providing an assortment of procedures that adequately cover the desired service region. If a patient requires one of the reallocated procedures, a suitable provider in close proximity could be identified and suggested to the patient. While the patient does not necessarily need to be served by the recommended provider, incentives such as a decreased copayment or priority scheduling might convince the patient to select the recommended provider.

Another way to utilize the analysis would be to create a list of providers with whom to form service contracts. These contracts could potentially benefit both the employer and the providers: employers could benefit from decreased cost per procedure, while providers could benefit from a guaranteed minimum volume of demanded procedures.

Additionally, patients could be empowered by the findings in the data by better understanding the costs associated with their health care. While patients are somewhat immunized against the complete costs of their health care via insurance, they nonetheless are conscious of their deductibles and copayments. Providing patients with additional insight into cost variations that exist in their area would allow them to play a more proactive role in making decisions related to their health care.

2.6 Future Directions and Conclusion

We have presented a data-driven approach to optimally selecting health care providers based on cost and convenience to patients. This approach is easy to implement and provides approximately 10% in expenditures for medical procedures. It is important to note that this estimate is based on conservative selection of parameters in our optimization approach. In practice, relaxing some of these parameters might yield opportunities for saving up to 15% in total health care expenditures. It also provides a starting point for the selection of a smaller collection of low-cost providers with whom an employer may negotiate for even better pricing.
Many of the procedures that the optimal solution has selected are diagnostic procedures, which do not vary substantially in quality from provider to provider. This, combined with the fact that for many patients to be reassigned in the data to a particular provider we must see high existing demand, suggests that the change in overall quality for the patient population would not differ too much.

Skyrocketing health care costs in the US are unsustainable and driven in part by highly variable pricing. Using real world historical claims data, we demonstrate that a reduction in the cost of medical procedures by approximately 10% can be achieved via relatively simple means. These cost savings can be easily implemented so that medical providers are within close proximity of the patient population; furthermore, because we utilize a mixed integer linear optimization model, it is very simple to add side constraints.

We view this framework as being widely applicable to all self-insuring employers in the US as a way of mitigating the costs associated with providing health care for employees. By utilizing historical claims data to identify and select medical providers, employers can protect themselves against rampant price variation in the currently opaque market for medical procedures. Ultimately, we hope that the implementation of this framework will contribute to the sustainability of health care costs in the US.
Chapter 3

Accept or Decline? An Analytics-Based Decision Tool for Kidney Offer Evaluation

3.1 Introduction

The current demand for kidney transplantation continues to outpace the supply; since 2002, the number of candidates on the waiting list has nearly doubled, from just over 50,000 to more than 96,000 by 2013 [Matas et al., 2015]. In contrast, living donation rates have decreased since 2002, and the deceased donation rate has not increased in recent years. To further compound the problem of organ supply and recipient demand, there is an unacceptably high deceased donor organ discard rate—up to 50% for high kidney donor profile index (KDPI) organs [Matas et al., 2015] and disproportionately high for increased infectious risk donors (IRD) when compared to non-IRD counterparts [Kucirka et al., 2009, Kucirka et al., 2011]. The decision to accept a deceased donor organ involves a complex series of well-intentioned “what if” calculations on the part of the accepting physician that rely on experience and intuition, instead of on data and patient outcomes.

The desire to provide patients with the highest quality organ holds potential to
become our Achilles heel. The desire to maximize the outcome for an individual patient can increase discard rates and overlooks the aggregate benefit achieved with an overall expansion of organ transplantation. It is at this crossroad that the physician’s semi-quantitative calculus and “gut” instinct come to the forefront and would benefit from a data-driven tool to assist in this complex decision-making algorithm.

We develop an analytics tool to assist physicians confronted with deceased-donor kidney acceptance decisions by utilizing a model based on patient-specific data regarding previous organ offers. In doing so, we seek to estimate an individual patient’s waiting time, in the context of a current active organ offer, until the time to the next offer for a kidney of a particular KDPI threshold or lower. If widely adopted, such a tool could not only provide reassurance for organ offer acceptance, but may also help improve the overall efficiency of kidney transplant allocation.

3.2 Data and Methods

We aim to address the following prediction problem: what is the probability of a patient being offered a deceased-donor kidney satisfying some KDPI threshold (or lower) within some time frame, given their individual characteristics? We consider different specifications with KDPI thresholds of 0.2, 0.4, and 0.6, and time frames of three, six, and twelve months. Because KDPI is an imperfect quality metric, we only consider offers not previously declined by more than 50 candidates in all three specifications. Nonetheless our methods can readily accommodate alternative KDPI thresholds, time frames, and previous offer decline cutoffs. Individual patient characteristics rely upon waitlist registry information and include Organ Procurement Organization (OPO) of listing, blood group, accumulated wait time, cPRA, age, DR antigens, and information regarding past offers (date, KDPI, and match run sequence number from previous offer(s)).
3.2.1 Data

We obtained waitlist and deceased donor information for the period April 1994 to September 2013 from the Organ Procurement and Transplantation Network (OPTN) Standard Transplant Analysis and Research dataset [2014 Annual Report]. We further obtained match run information for deceased donor kidneys that were eventually accepted by a patient during 5 March 2007 to 30 June 2013 from the OPTN Potential Transplant Recipient (PTR) dataset. For the period that our PTR data spans, we retrieved 38806 match runs and their associated donor information; in this data, 535 of the match were directed donations or assigned for multi-organ transplants, and we remove them. For the same period, we retrieved waitlist information for the 287283 patients who were active for at least some fraction of it.

3.2.2 Methods for Wait time Prediction

We address the prediction problem using supervised learning methods, that is, methods that are trained on historical data [Bertsimas et al., 2016]. Specifically, we use random forest, a well-studied method in the machine learning literature that has been widely applied to prediction problems [Bertsimas et al., 2016, Breiman, 2001, Liaw and Wiener, 2002]. For each specification, our random forest models predict the probability with which a patient would be offered a kidney within the specified time frame (the dependent variables), given their characteristics and offer histories (the independent variables). So that these models can make such predictions, we first train them using historical observations of independent variables and their associated dependent variables. Once trained, the models then predict as output the dependent variables, given (potentially previously unseen) observations of the independent variables.

3.2.3 Observations and Independent Variables

An observation corresponds to a patient registered at a transplant center in one of the ten most populous OPOs and a random date within the time period of our
PTR dataset. For each particular patient, we sample eight random dates—therefore generating eight observations—for every year accrued on the waitlist.

For each generated observation, we compute the dependent variable as 1 (0) if the patient was (not) offered a kidney of specified KDPI within the specified time frame from the observation date. We censor observations for which patients were inactive for more than 90% of the follow-up time frame or became inactive as a result of accepting a kidney of KDPI higher than the specified threshold during the follow-up time frame. Consequently, the number of observations per patient is essentially proportional to the time they had been active on the waitlist.

We compute the following independent variables for each observation:

1. The patient’s OPO;
2. The patient’s blood group;
3. The patient’s accumulated wait time on the waitlist;
4. The expected number of DR mismatches that the patient would have with the next kidney to be offered;
5. An indicator variable of whether or not the patient was pediatric;
6. An indicator variable of whether or not the patient was sensitized (cPRA > 0.8);
7. The number of kidneys of KDPI ≤ 0.4 offered to the patient prior to the observation date with organ sequence number being at most 500, divided by the number of days the patient was active prior to the observation date;
8. The same as previous, except for kidneys of quality 0.4 < KDPI ≤ 0.7;
9. The same as previous, except for kidneys of quality KDPI > 0.7;
10. The patient’s average organ sequence number in the match run for offers they received prior to the observation date;
11. The mean number of DR mismatches with kidneys offered to the patient prior to the observation date;

12. The average KDPI of kidneys offered to the patient prior to the observation date;

13. The patient’s current cPRA value at the observation date;

14. A factor variable indicating the month of the observation.

Each independent variable adds predictive ability to the model. Because patients belonging to different OPOs and blood groups are subject to different rates of waitlist congestion and donation rates, variables 1-2 provide context for the remaining variables. Variables 3-6 contribute directly to the points system that is used to rank patients in match runs. To compute variable 4, we compare the patient’s DR antigens with that of all previous donor kidneys (from the same OPO) prior to the observation date and then calculate the average number of mismatches. In essence, this captures the points that the patient was expected to receive due to DR antigen matching. Variables 7-9 measure the frequency of receiving offers prior to the observation date, for each of the quality categories considered. Because including all organ offers greatly reduces predictive power, we consider only offers with sequence number at most 500; we make this determination via validation. Additionally, as candidates are not offered kidneys during periods of inactivity, variables 7-9 self-account for the time that patient had been inactive. Variables 10-12 provide an indication of the patient’s ranking and of the quality of the organs offered at previous match runs. Variable 13 informs the model with an indication of how many procured kidneys would be compatible with the patient. Variable 14 captures fluctuations that were observed between the number of deceased donors in any given month.

It is worth noting that, with the exception of the last variable controlling for month of the year, all of the other variables are specific to the patient, thus allowing the model to make personalized predictions for waitlist candidates.
3.2.4 Performance Evaluation

We split observations spanning the entire time period of available match-run data into training, validation, and testing sets in chronological order, a standard performance evaluation procedure in the data analytics literature [Bertsimas et al., 2016]. For example, for the twelve-month time frame specification, the training set includes all observations whose dates fell between 1 May 2007 and 30 April 2008. The validation set includes observations dated between 1 May 2009 and 30 April 2010, while the testing set includes observations dated between 1 May 2011 and 30 April 2012. There is a twelve-month gap between each of these cut-off dates, since the dependent variable captures information up to twelve months after the observation date.

Out-of-sample Area Under the Receiver Operating Characteristic Curve (AUC)

We fit a random forest model on the training set and compute the out-of-sample area under the receiver operating characteristic curve (AUC) for the validation set. AUC is a standard metric for determining a model’s ability to distinguish between two outcomes in a binary prediction model. In this setting, for example, a model with an out-of-sample AUC value of 0.60 can correctly distinguish between a randomly selected negative example and a randomly selected positive example 60% of the time.

We vary the number of trees and the number of variables per tree as the parameters to the random forest. We select the parameter values that yielded the highest AUC for the validation set.

To finalize the model, we re-train it on combined training and validation sets for the selected parameter values. To simulate real-world out-of-sample testing procedures, we only evaluate performance on the testing dataset using the finalized model.

Accuracy Within Clusters of Similar Patients

To evaluate the accuracy of the model, we evaluate out-of-sample predictions over clusters of “similar” patients as follows. First, we cluster observations in the testing set
by OPO, blood group, the number of years the patients had accrued on the waitlist, and their DR antigens. Second, for each cluster and each dependent variable, we compare the variable’s average value over the cluster that our model predicted with the actual average value in the testing set. Put differently, this allows the comparison of the average number of candidates that our model predicted to receive an offer of the specified KDPI within the specified time frame with the average number of candidates that actually received such an offer in practice.

Creating Datasets with Rolling Horizons to Assess the Adaptability of the Method

To assess the accuracy of our models as the amount of data collected increases, we create prediction models on subsets of the data starting with just the first three years of data, gradually increasing to the entire dataset.

Figure 3-1A) illustrates how three years of data can be used for the three-month time frame specification. We use a three-month buffer after each of the training, validation, and testing sets to investigate predictions within three months. When the amount of data available is increased to four years, we can then splice the data into training/validation/testing sets in multiple ways as in Figures 3-1B) and 3-1C). For a given amount of data and a time frame specification, we create all possible splits of training/validation/testing sets that are feasible, limiting both validation and testing sets to either six months, nine months, or one year. For each possible split, we train prediction models and obtain their out-of-sample AUC values.

Disclaimer

PTR data were obtained under IRB approval (#F23797-101) from the Harvard University Committee on the Use of Human Subjects in Research. All identifiers were removed upon data receipt for the purposes of this study.

This study used data from the Organ Procurement and Transplantation Network (OPTN). The OPTN data system includes data on all donor, wait-listed candidates, and transplant recipients in the US, submitted by the members of the Organ Procure-
3.3 Results

3.3.1 AUC

The model enjoys excellent out-of-sample performance, with AUC values of 0.86, 0.88, and 0.87 when predicting the probability of receiving an offer with KDPI ≤ 0.2 in three, six, and twelve months, respectively. Table 3.1 also reports out-of-sample AUC results for KDPI thresholds of 0.4 and 0.6 for all time frames we consider. We observe that the model’s performance remains consistent for all thresholds considered.

To illustrate which variables contribute the most to the random forest model’s predictive power, we present variable importance plots in Figures 3-2—3-4 for the KDPI ≤ 0.4 specification. We observe that average organ sequence number is the most important variable for each of the three time frames. Years on the waitlist,
### Table 3.1: Out-of-sample AUC values for the random forest model for the three-, six-, twelve-month time frames with KDPI thresholds of 0.2, 0.4, and 0.6. We observe that the results are robust to changes in time frame as well as the desired offer quality.

<table>
<thead>
<tr>
<th>KDPI Threshold</th>
<th>3 Months</th>
<th>6 Months</th>
<th>12 Months</th>
</tr>
</thead>
<tbody>
<tr>
<td>KDPI ≤ 0.2</td>
<td>0.862</td>
<td>0.881</td>
<td>0.866</td>
</tr>
<tr>
<td>KDPI ≤ 0.4</td>
<td>0.878</td>
<td>0.881</td>
<td>0.870</td>
</tr>
<tr>
<td>KDPI ≤ 0.6</td>
<td>0.880</td>
<td>0.880</td>
<td>0.881</td>
</tr>
</tbody>
</table>

We observe that average organ sequence number for previous offers, years on the waitlist, intensity of previous kidney offers, and the month of prediction are the five most important variables for the 3-month time frame.

#### 3.3.2 Accuracy Within Clusters of Similar Patients

Figure 3-5 presents an accuracy comparison between actual and predicted fractions of patients for blood group O in California (CAOP) receiving offers of KDPI ≤ 0.4 within six months, broken down by clusters based on years accumulated on the waitlist. For example, for the cluster of patients who have accumulated 5 years of wait time, our model predicts that 2.6% of them will receive a KDPI ≤ 0.4 kidney offer within six
Figure 3-3: Variable importance plot for the six-month, KDPI ≤ 0.4 specification. We observe that the top five most important variables remain the same as for the 3-month time frame, with a slight change in ordering. OPO increases in importance when compared to the 3-month time frame.

Figure 3-4: Variable importance plot for the twelve-month, KDPI ≤ 0.4 specification. We observe that OPO has again increased in importance and is now the second most important variable, and has displaced intensity of KDPI ≤ 0.4 kidneys from the top five most important variables. Average organ sequence number for previous offers remains the most important variable throughout.
months, whereas the actual fraction is 3.9%. Figure 3-6 presents a similar comparison for all of the blood groups within CAOP.

Predicted vs. Actual Probability of KDPI ≤ 0.4 Offer for Type O in CAOP in 6 Months by Years on Waitlist

Figure 3-5: Predicted versus actual probabilities of a KDPI ≤ 0.4 offer within six months for Type O in California (CAOP).

Figure 3-7 demonstrates a similar comparison for patients in New York (NYRT) clustered by both wait time and DR antigen profile. Patients with more compatible DR antigen profiles tend to have higher chances of receiving a KDPI ≤ 0.4, both in practice and in our model predictions, demonstrating that the models provide reasonable accuracy when accounting for a variety of specific patient characteristics. For the example depicted in Figure 3-7, the most compatible 33% of patients (i.e., patients with fewer than 1.61 expected DR mismatches) are predicted to be 4.9% more likely to receive a KDPI ≤ 0.4 organ in year 6 on the waitlist compared to the least compatible 33% of patients (i.e., patients with more than 1.71 expected DR mismatches), whereas the actual difference is approximately 5.4%.
Figure 3-6: Predicted versus actual probabilities of a KDPI ≤ 0.4 offer within six months in California (CAOP). Our models correctly capture the fact that different blood groups are subject to different waiting times.

### Example of Intended Usage

A deceased donor kidney with KDPI of 0.55 is offered to patient J.S. who has accumulated five years and four months of waiting time. J.S.’s blood group is O, cPRA is 5%, and DR antigens are 7 and 13. Two months ago, J.S. was offered a kidney with KDPI of 0.83, and on that match run had sequence number 153. What is the probability that J.S. will be offered a “high-quality” kidney (KDPI ≤ 0.4) within the next six months?

Table 3.2 demonstrates the model’s predictions, for several different OPOs, of J.S.’s probabilities of receiving a KDPI ≤ 0.4 offer within the next six months. We also report the actual probabilities that patients similar to J.S. received such offers in practice. For example, in Pennsylvania (PADV), 19.0—24.2% of patients similar
Figure 3-7: Predicted versus actual probabilities of a KDPI ≤ 0.4 offer for patients of blood group O in New York (NYRT) within six months, depending on whether their DR antigens are in the top 33% of compatibility (< 1.6 expected DR mismatches) or lowest 33% of compatibility (> 1.7 expected DR mismatches). Notice that patients with DR antigens that have historically had fewer mismatches have higher predicted and actual probabilities.

to J.S. receive KDPI ≤ 0.4 offers within six months; the model’s prediction for J.S. is 22.7%.

<table>
<thead>
<tr>
<th>Region</th>
<th>Actual Probability</th>
<th>Model’s Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>California (CAOP)</td>
<td>3.4 – 3.7%</td>
<td>3.8%</td>
</tr>
<tr>
<td>Illinois (ILIP)</td>
<td>8.3 – 15.7%</td>
<td>11.7%</td>
</tr>
<tr>
<td>New York (NYRT)</td>
<td>4.2 – 8.6%</td>
<td>7.9%</td>
</tr>
<tr>
<td>Pennsylvania (PADV)</td>
<td>19.0 – 24.2%</td>
<td>22.7%</td>
</tr>
<tr>
<td>Texas (TXGC)</td>
<td>23.3 – 26.2%</td>
<td>30.7%</td>
</tr>
</tbody>
</table>

Table 3.2: Predicted versus actual probabilities that patient J.S. will be offered a kidney of KDPI ≤ 0.4 within the next six months.
3.3.4 On the Adaptability of the Prediction Method

The OPTN kidney allocation system (KAS) was revised in December 2014. Due to the new rules instituted and the potential for even further modifications, it is reasonable to expect that new prediction models will have to be trained as data is accrued under any new allocation system, in order to achieve high-quality out-of-sample results. However, the extent of the data required in order to achieve high-quality AUC results remains to be determined. To further investigate this, we created a rolling horizon starting with 3 years of data and evolving to the complete dataset range of 6 years, with observed changes in AUC under an increasing amount of available data. We detail the construction of these rolling horizons in the methods section.

Figure 3-8 depicts the observed evolution of AUC for predicting offers of KDPI ≤ 0.4 within three, six, or twelve months, demonstrating that with additional data, AUC values experiences a slight increase before leveling off.

Figure 3-8: AUC value evolution over increasing amounts of data for each time frame considered. Each plot is for a different time frame (three, six, or twelve months), and the horizontal axis denotes the total number of years of data. Note that each point is averaged over results obtained from splicing the data into training/validation/testing in various ways, given a particular rolling horizon.
3.4 Discussion

Transplant practitioners and waitlisted candidates are confronted with the crucial decision of whether to accept or decline an offered organ, often in the setting of a limited time window during which to make a decision. The current paradigm relies on a practitioner’s experienced-based approach, which lacks scientific rigor and is subject to unreliable and irreproducible decision-making. We developed an analytical tool that can not only assist practitioners in their organ acceptance decisions at these critical moments, but also serves as an educational tool for candidates awaiting kidney transplantation. In defining the future organ offer landscape in a patient-specific format, we hope to not only provide transplant practitioners the ability to achieve expedited, evidence-based, decision-making for organ selection, but also to provide an interactive educational tool for transplant candidates to further their understanding of an additional aspect of the risk/benefit ratio associated with organ offer acceptance—specifically the factor of additional waiting time.

The problem of deciding whether to accept or decline a particular offered kidney, so as to maximize a patient’s overall outcome and maintain efficiency in the allocation system, can be thought of as an optimal stopping problem: how long should a patient wait prior to accepting an offered kidney that best suits their needs? By developing data-driven predictive analytics models that provide information about the probability of being offered high-quality organs in the future, we are closer to solving the optimal kidney acceptance problem. Indeed, the next extension of these predictions will rely on developing dynamic optimization models capable of assessing the accept/decline tradeoff by incorporating data regarding the patient’s current condition on dialysis. Ultimate application will require an initial assessment in a controlled prospective fashion to ascertain the utility and benefit of a decision support tool for kidney offer acceptance.

From a more global perspective, properly contextualizing the probability of receiving a future higher quality kidney offer, in comparison to a candidate’s current offer, will facilitate the organ offer decision-making process. Although this conceivably
holds potential to allow for a more efficient process of donor and recipient matching and ultimately contribute to fewer discarded kidneys, this remains to be determined and will require application of the predictive model in a prospective fashion.

The decision to proceed with organ acceptance for transplantation incorporates decision-making based on a multitude of factors, including: donor quality, recipient condition, the incompatibility between the two, as well as center-specific or practitioner-specific influences. Importantly, the calculus associated with accepting a deceased-donor organ offer revolves around the quality of the organ at hand compared to the anticipated organ quality of the next offer in the context of a patient’s waiting time and co-morbidities. However, the various weight applied to these individual factors likely varies based on which organ is being transplanted. Volk et al. recently demonstrated an organ support tool to assist practitioners with organ offers for candidates waitlisted for liver transplantation with the hopes of improving patient survival in liver transplantation [Volk et al., 2016]. The decision to proceed with a kidney offer likely differs from that of other solid organ transplants (i.e., liver, heart, or lung), for which a more immediate “life-or-death” situation guides the decision-making process. Indeed, for these non-renal organs, the immediacy of the candidate’s demise while awaiting the next “higher quality” offer likely holds stronger influence in the decision-making process for offer acceptance when compared to that for kidney transplant candidates who can maintain stability (although not indefinitely) through renal replacement therapy. In fact, we address a more pertinent problem for liver transplantation in the next chapter. Given the kidney transplant candidates’ option for dialysis, and reduced risk of demise precluding transplant while waiting for the “next” offer, we hypothesize that a decision support tool may have its greatest impact in renal transplantation.

Our dataset included deceased donor kidneys between the years of 2007 and 2013. On 4 December 2014, OPTN introduced a new kidney allocation system. Although the data presented herein, and thus the results, were based on the prior allocation system, the new system introduced resulted in a more uniform allocation process across OPOs, diminishing the effect of regional variances in allocation. Our mod-
els are already tailored to account for OPO-specific differences that were observed previously; thus, decreasing the variation among OPOs would likely strengthen the predictive power of our models. In spite of these OPO variances within our dataset, the current results show promise for extending these methods for predictions under the new allocation rules once enough data has been accrued. Indeed, our results on the adaptability of the prediction model demonstrate that under the new allocation system, sufficient data has yet to be collected to implement this prediction method at three, six, and twelve months. It is anticipated that the combination of further accrual of data under the new system, with a reduction in variability between OPOs, may allow for improved AUC results compared to our current findings.

It is important to note that our current model is proof of principle of the power for an analytics decision tool to enhance organ offer decision making. It is thus anticipated that with time, the datasets generated from the new allocation schemes will be applied in the same context. Indeed, the current models need to be trained on historical data only once; thereafter, the process of making predictions for specific patients can be computed almost instantaneously.

The dataset examined consists only of kidneys that were eventually accepted by a candidate, and thus did not allow for examination of kidneys that were discarded; the latter leads to a degree of biasing in our dataset. Incorporation of discard data will allow better predictions about the probability with which patients might be offered lower quality kidneys, and would thus strengthen the overall prediction capabilities of the model. It should also be noted that although the resultant models utilizing random forests are not easily interpretable, when we applied the more interpretable classification trees pioneered by Breiman et al. [Breiman et al., 1984], we observed significantly weaker AUC when compared to random forest, as noted in Table 3.3. We also experimented with using logistic regression as our prediction model and found its AUC to be weaker than the AUC of the random forest model, as noted in Table 3.4. Moreover, we also found that the logistic regression was significantly less accurate for individual predictions, which we detail in Table 3.5.
<table>
<thead>
<tr>
<th>Time Frame</th>
<th>Random Forest</th>
<th>Classification Tree</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 Months</td>
<td>0.878</td>
<td>0.814</td>
</tr>
<tr>
<td>6 Months</td>
<td>0.881</td>
<td>0.809</td>
</tr>
<tr>
<td>12 Months</td>
<td>0.870</td>
<td>0.792</td>
</tr>
</tbody>
</table>

Table 3.3: Out-of-sample AUC values for random forest and classification tree.

<table>
<thead>
<tr>
<th>Time Frame</th>
<th>Random Forest</th>
<th>Logistic Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 Months</td>
<td>0.878</td>
<td>0.852</td>
</tr>
<tr>
<td>6 Months</td>
<td>0.881</td>
<td>0.851</td>
</tr>
<tr>
<td>12 Months</td>
<td>0.870</td>
<td>0.841</td>
</tr>
</tbody>
</table>

Table 3.4: Out-of-sample AUC values for random forest and logistic regression.

<table>
<thead>
<tr>
<th>Region</th>
<th>Actual Prob.</th>
<th>Random Forest</th>
<th>Logistic Reg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>California (CAOP)</td>
<td>3.4 – 3.7%</td>
<td>3.8%</td>
<td>15.0%</td>
</tr>
<tr>
<td>Illinois (ILIP)</td>
<td>8.3 – 15.7%</td>
<td>11.7%</td>
<td>24.1%</td>
</tr>
<tr>
<td>New York (NYRT)</td>
<td>4.2 – 8.6%</td>
<td>7.9%</td>
<td>21.4%</td>
</tr>
<tr>
<td>Pennsylvania (PADV)</td>
<td>19.0 – 24.2%</td>
<td>22.7%</td>
<td>31.4%</td>
</tr>
<tr>
<td>Texas (TXGC)</td>
<td>23.3 – 26.2%</td>
<td>30.7%</td>
<td>51.1%</td>
</tr>
</tbody>
</table>

Table 3.5: Predicted versus actual probabilities that patient J.S. will be offered a kidney of KDPI ≤ 0.4 within the next six months for random forest and logistic regression. Notice that the logistic regression consistently overestimates the probability of an offer.

Application of methods from the data analytics and machine learning literature has allowed for the development of predictive models to provide personalized probabilities of receiving kidneys of specified quality within time frames of three, six, and twelve months. We demonstrated that these models have strong predictive ability out-of-sample, and thus can be used in real-world scenarios. Importantly, our models are strictly data-driven and rely on minimal assumptions. By informing physicians and transplant candidates with these probabilities, we hope to progress from intuition-based decision-making, to actionable insight distilled from data.
Chapter 4

Who Deserves Priority? Predicting Adverse Events on the Liver Transplant Waitlist

4.1 Introduction

Unlike the kidney transplantation waitlist where patients are ranked predominantly based on the amount of wait time they have accrued on the waitlist, end-stage liver disease patients on the liver transplantation waitlist are prioritized based on medical urgency. This is due to the lack of a comparable procedure to dialysis for livers; extremely severe cases of liver failure require expedient transplantation so as to maximize the chances of survival. As a result, one of the most crucial challenges in liver transplantation involves accurately determining a particular patient’s likelihood of death or unsuitability for transplantation within the near future so that the limited supply of donated livers can be allocated to maximize the benefit from transplantation.

Although the current organ allocation process depends on the Model for End-Stage Liver Disease (MELD) score to rank disease severity—and, as a result, priority for receiving a liver—the MELD score has several weaknesses that have led to in-
equitable and undesirable patient outcomes. In particular, the MELD score has undergone multiple revisions and tweaks to accommodate the important subpopulation of liver cancer patients [Elwir and Lake, 2016]. Even with these adjustments, it does not adequately capture the difference between cancer and non-cancer patients; more importantly, the various proposed adjustments give an unfair advantage to cancer patients at the expense of other patients.

In this chapter, we discuss the advantages and disadvantages of using the MELD score and propose a new data-driven analytics-based method to predict the rate of adverse events (either death or unsuitability for transplant) within three months for patients on the liver transplantation waitlist. Our model not only outperforms the currently implemented MELD policy in accurately predicting either death or unsuitability for transplant, but also addresses several of its shortcomings with respect to its inability to generalize to cancer patients. More importantly, our model increases the number adverse events averted over the cancer-adjusted MELD score in simulation by approximately 25.5% for all patients and 2.8% for cancer patients. It is our aspiration that our method can contribute to more equitable and optimal patient outcomes.

4.2 History and Current Practice: Model for End-Stage Liver Disease (MELD)

The United Network for Organ Sharing (UNOS) adopted the original iteration of MELD in 2002 [Bernardi et al., 2011]; at the time, this represented a major change in the way organs were allocated to end-stage liver disease patients. To better understand the reasons for adopting MELD, it is helpful to delve into how organ allocation was decided before its introduction.
4.2.1 The Pre-MELD Era

Prior to the use of MELD in ranking patients on the liver waitlist, the allocation policy was similar to that of the kidney waitlist: patients who had accrued the most time on the waitlist were given higher priority. However, this policy led to highly unequal outcomes. Because disease progression differs from patient to patient, it was often the case that livers were allocated to patients who had listed earlier, at the expense of patients who were more ill. Several transplantation centers observed this unequal allocation and attempted to correct for it by stratifying patients into varying care intensity statuses, such as “chance of survival less than 24 hours” (status 1) and “intensive care unit” (status 2), to “home management” (status 4) [Bernardi et al., 2011]. Under this new criteria, transplantation centers first considered patient status before referring to accrued time on the waitlist when deciding to which patients livers would be offered.

The immediate precursor to the MELD score was the Child-Turcotte-Pugh (CTP) score. This score, originally designed to predict patient outcomes for port-caval shunt surgery or esophageal transection for bleeding varices, was the first time clinical data and laboratory variables were used to determine patient eligibility for liver offers [Bernardi et al., 2011]. Although the CTP score was a step in the right direction, it could only group patients into four buckets of disease severity; because it was unable to provide personalized mortality predictions for patients, the CTP score relied predominantly on patient waiting time as a tiebreaker.

4.2.2 MELD

Advantages and Disadvantages

The first iteration of MELD was pioneered in 2000 [Malinchoc et al., 2000]. Although originally designed to predict the outcome of transjugular porto-systemic shunt in hypertensive and cirrhotic patients, the MELD score demonstrated impressive performance in predicting the three-month mortality rate for patients, with the literature reporting AUC values of approximately 0.83 [Wiesner et al., 2003].
One of the advantages of the MELD score in comparison to its predecessors was that it relied on only three laboratory values: serum creatinine, serum bilirubin, and international normalized ratio (INR) for prothrombin time. This ensured that only objective measures were used in determining disease severity. MELD also improved upon other existing methods for allocation by introducing for the first time a continuous metric by which to rank patients. The original unadjusted MELD score was computed as follows:

$$9.57 \log(\text{creatinine mg/dL}) + 3.78 \log(\text{bilirubin mg/dL}) + 11.20 \log(\text{INR}) + 6.43,$$

where lab values have a lower limit of 1.0, and creatinine is capped at 4.0 mg/dL [MELD Calculator, 2009].

After adopting the MELD score to prioritize patients on the liver waitlist, UNOS observed a significant decrease in median waiting time to liver transplantation as well as a lowered death rate for patients on the waitlist in comparison to the pre-MELD era by approximately 10-20% [Bernardi et al., 2011]. The exact effect was difficult to measure, in part due to an increase in the organ procurement rate which coincided with the introduction of MELD [Bernardi et al., 2011].

One of the biggest drawbacks of the MELD score is that it does not adequately model mortality for patients with hepatocellular carcinoma (HCC), a particular type of liver cancer [Bernardi et al., 2011]. These patients, who account for approximately 10-20% of registrants on the liver transplantation waitlist, often exhibit good hepatic function over long periods of time, leading to values of creatinine, bilirubin, and INR that yield a low MELD score. As a result, the MELD score severely underestimates the mortality rate for this important patient subpopulation. Furthermore, HCC patients awaiting liver transplantation are subject to the additional complication that their disease may progress beyond the point where liver transplantation is suitable. All of these deficiencies are well-noted in the literature, with Bernardi et al. noting:

“To overcome this potential shortcoming, modifications consisting in adding points to the calculated MELD score have been proposed, some of which
are widely employed...However, these adjustments were created arbitrarily, often without scientific evidence, with the purpose of avoiding an advantageous or disadvantageous prioritization vis-a-vis cirrhotic patients without these conditions for whom the MELD score had been validated.”

[Bernardi et al., 2011]

Modifications

The difficulty in handling HCC patients mentioned above led to a series of heuristic and, as described in the literature, arbitrary modifications to the original iteration of the MELD score in recent years. In particular, to compensate for the unadjusted MELD score’s underestimation of disease severity for HCC patients, the medical community instituted an artificially boosted MELD score for HCC patients (which we denote HCC-adjusted MELD), depending on their wait time and disease progression [Elwir and Lake, 2016].

This series of complex adjustments to MELD are not well understood, especially since many of the thresholds were selected based on intuition. Indeed, retrospective analyses have been conducted after the modifications were implemented to ensure that the changes had a desirable impact on organ allocation [Elwir and Lake, 2016].

At the same time, in 2006 Biggins et al. observed that adding sodium values into the computation of the MELD score added predictive power for patients experiencing hyponatremia [Biggins et al., 2006]. After several years of discussion and amidst all of the other modifications to using MELD for allocation, UNOS approved the new MELD-Na score in June 2014; the new MELD-Na score was officially instituted for implementation purposes in January 2016 [Elwir and Lake, 2016].

4.2.3 Beyond MELD: An Integrated Model

Although initially MELD held great promise by decreasing mortality rates on the liver transplantation waitlist by shifting the focus from wait time to mortality, the shortcoming with respect to inequality of HCC patients necessitated the implementation
of a series of reactive modifications. Given that these modifications were based on intuition as opposed to science, it is appropriate for us to consider a new, integrated model that measures the benefits of transplantation for all patients.

### 4.3 Data and Methods

We aim to address the following prediction problem: what is the probability that a patient will either die or become unsuitable for liver transplantation within the next three months, given their individual characteristics? Although we consider three months for this particular setting so that we can make a direct comparison to both unadjusted MELD and HCC-adjusted MELD, it is important to note that our methods can readily accommodate alternative time frames.

#### 4.3.1 Data

We obtained waitlist, deceased donor, transplant, and follow-up information for the period 1 January 2002 to 5 September 2016 from the Organ Procurement and Transplantation Network (OPTN) Standard Transplant Analysis and Research (STAR) dataset [2014 Annual Report]. For our waitlist data, each row represents a patient receiving a periodic check-in visit after registering for the liver waitlist. For instance, since patients must have their laboratory values (creatinine, bilirubin, INR, sodium) taken at regular intervals to determine their MELD score, a patient may generate multiple rows as they accrue time on the waitlist.

#### 4.3.2 Methods for Predicting Benefit of Transplantation

An important distinction between the prediction problem we address and the problems on which MELD and its variants were trained is that we consider both mortality and unsuitability for transplant. Thus, from the outset, our methods are designed to handle and account for HCC patients implicitly, with no need to tack on adjustments ex-post.
Given our data and problem specification, we can utilize a variety of supervised learning models, that is, data analytics methods that are trained on historical data [Bertsimas et al., 2016]. As in the previous chapter, we use random forest. We first train our random forest model using historical observations of independent variables and their associated dependent variables. Once trained, the model then predicts as output the dependent variables, given (potentially previously unseen) observations of the independent variables.

4.3.3 Observations and Independent Variables

When curating our data for the model, each observation corresponds to a patient at the time of their check-in visit. The STAR dataset includes information on the patient’s laboratory values at that particular point in time.

For each observation, we compute the dependent variable as 1 (0) if the patient experienced (did not experience) an adverse event, including death or unsuitability for transplant, during a follow-up period of three months from the check-in date. We then computed the independent variables as follows:

1. Whether or not the candidate has HCC (is a cancer patient);
2. The candidate’s recorded albumin level;
3. The candidate’s recorded ascites status;
4. The candidate’s encephalopathy status;
5. The candidate’s serum bilirubin;
6. The candidate’s serum creatinine;
7. The candidate’s INR or prothrombin time;
8. The candidate’s serum sodium level;
9. The number of times the candidate received dialysis in the last week;
10. The candidate’s albumin level at previous check-in;

11. The candidate’s ascites status at previous check-in;

12. The candidate’s encephalopathy status at previous check-in;

13. The candidate’s serum bilirubin at previous check-in;

14. The candidate’s serum creatinine at previous check-in;

15. The candidate’s INR at previous check-in;

16. The candidate’s serum sodium at previous check-in;

17. The change in the number of times the candidate received dialysis since previous check-in;

18. The change in the candidate’s bilirubin level since previous check-in;

19. The change in the candidate’s creatinine level since previous check-in;

20. The change in the candidate’s INR since previous check-in;

21. The change in the candidate’s albumin level since previous check-in;

22. The change in the candidate’s sodium level since previous check-in;

23. The number of years the candidate has accrued on the waitlist;

24. The candidate’s age in years.

Each independent variable adds predictive ability to the model. Variable 1 informs the model as to whether or not the patient has cancer; cancer patients have different disease trajectories that affect their mortality rate. Variables 2-4 are clinical measurements that contributed to the CTP score, which was the precursor to MELD. Variables 5-8 are the raw clinical measurements that are used to determine the MELD score. Variable 9 indicates the patient’s current dialysis status, which affects the degree to which the patient’s creatinine reading influences the MELD score. Variables
10-17 capture information about the most recent check-up prior to the patient’s current health status update and informs the model as to how the patient’s condition has changed over time. These variables are particularly important, as they help to capture MELD spikes, or sudden sharp changes in a patient’s MELD score. Under the current allocation policy, patients undergoing these spikes are highly disadvantaged [Massie et al., 2015]. Variables 18-22 further capture the patient’s health trajectory. Variable 23 contributes information as to how long the patient has been waiting; this variable potentially captures important information about a patient’s cancer status, since patients with cancer often are asymptomatic and do not register for the waitlist early enough. Variable 24 informs the model of the patient’s age, which is an important factor in predicting mortality.

4.3.4 Performance Evaluation

For consistency, we only considered observations that were taken after MELD was initially implemented. We split all observations from 2002 and onward into training, validation, and testing sets in chronological order, a standard performance evaluation procedure in the data analytics literature [Bertsimas et al., 2016]. For training and validation, we considered all data before 1 October 2014. The testing set consisted of all observations beginning on or after 1 January 2015. There was a three-month gap between the last observation in the training and validation sets and the first observation in the testing set since the dependent variable captured information up to three months after the observation date.

Out-of-sample AUC values

We fit a random forest model on the training set and computed the out-of-sample AUC value for the validation set. We vary the number of trees and the number of variables per tree as the parameters to the random forest, and select the ones that yield the highest AUC for the validation set. We finalize our model by re-training it on combined training and validation sets for the selected parameter values. To
simulate real-world out-of-sample testing procedures, we only evaluate performance on the testing dataset using the finalized model.

**Adverse Events Averted by Each Model**

In addition to out-of-sample AUC performance, we are interested in how well our random forest model ranks the sickest patients by probability of death or unsuitability for transplant within three months. Because deceased donor livers are a scarce resource, it is crucial that the score used to rank disease severity and, correspondingly, risk of experiencing an adverse event is correct for the patients most at risk. Indeed, inaccurately predicting the risk adverse events for the sickest patients is a life-or-death decision for the patients most in need of donor organs: depending on the scoring metric used, patients with overestimated risk may potentially knock other patients with underestimated risk further down the waitlist.

To compare the performance of our random forest model with the unadjusted MELD and HCC-adjusted MELD (i.e., the current score used in practice) scores in the desired use case, we create a simulation environment in which we consider sample paths of liver arrivals that were realized in the historical data, assign organs to the sickest patients as determined by each ranking method, and compute the number of adverse events that each scoring metric averted. Given a particular sample path of liver arrivals, we then run a simulation to determine the outcome—that is, the number of adverse events prevented—using unadjusted MELD, HCC-adjusted MELD and random forest as the scoring metrics.

It is critical to note that the number of adverse events averted will not be the same as the number of livers procured. A scoring metric with perfect foresight would be able to prevent the maximum number of adverse events, subject to liver availability. On the other hand, a poor scoring metric may assign livers to patients that are in good health and do not experience adverse events in the data.

Because organ procurement and allocation is, for the most part, conducted on the regional level, we have separated out each of our simulations by service region. For each service region, we run a simulation of the date range 1 January 2015 to 31
December 2015. Note that this date range is out-of-sample for our random forest model.

Our simulation is conducted as follows: we begin with all patients in that region with any status updates up to three months before 1 January 2015. For each day of the simulation, we run the following steps:

1. We update the most recent status obtained for each patient on the waitlist. Status updates are handled as follows:

   - Patients on the waitlist who experience an adverse event (either death or inability to receive transplant) are removed from the waitlist.
   - Patients who receive a transplant in the actual data within three months ahead of the date being considered are removed. This is due to the fact that these patients’ three-month survival rates are censored.
   - Patients on the waitlist who have a health status update have their mortality prediction updated.

2. Given this updated waitlist, we then take the number of organ arrivals (as realized on that day in previous years) and assign the sickest patients who are still on the waitlist—as determined by the particular model chosen—for transplant. These patients are removed from the waitlist.

For our simulations, we utilize liver procurement sample paths from years 2010-2014. This allows us to capture seasonality effects in organ procurement rates. For each of these five sample paths, we run the simulation with unadjusted MELD, HCC-adjusted MELD, and random forest. For each simulation trial, we compute the number of patients who were assigned a transplant by each method before the patients’ actual date of death or removal due to unsuitability for transplant. For each patient who was assigned a transplant before their actual date of removal from the waitlist, we consider this adverse event to be “averted”. The outcome of the simulation is thus the number of adverse events averted by each of the models. Note that it is possible
in our simulation for transplants to be assigned to patients who did not die, become unsuitable for transplant, or receive a transplantation within three months in the actual data. In this case, we do not consider the transplant to have averted an adverse event.

4.4 Results

4.4.1 AUC

The random forest model, with an out-of-sample AUC value of 0.800, considerably outperforms both unadjusted MELD score’s AUC of 0.754 and HCC-adjusted MELD’s AUC of 0.708 when predicting a patient’s three-month probability of experiencing an adverse event of either death or unsuitability for transplant. It is important to note that these AUC values for MELD are lower than in the literature, due to the fact that our prediction problem includes unsuitability for transplant in addition to mortality. To illustrate which variables contributed the most to the random forest model’s predictive power, we present the variable importance plot in Figure 4-1.

4.4.2 Adverse Events Averted by Each Model

Since we simulated each service region separately, our study of adverse events averted by each model is broken down by region. We present a map specifying which states belong to each particular region in Figure 4-2. Alaska and Hawaii, though not shown in the figure, belong to region 6.

Table 4.1 presents the mean number of adverse events that are averted for all patients in simulation by random forest, unadjusted MELD, and HCC-adjusted MELD for the five sample paths of organ arrivals obtained from previous years. In every region, the random forest metric significantly outperforms both unadjusted MELD and HCC-adjusted MELD in averted adverse events. For instance, in service region 5 the random forest model averts an additional 45 adverse events over unadjusted MELD and an additional 104 adverse events over HCC-adjusted MELD.
Figure 4-1: Variable importance plot for the 3-month adverse event prediction random forest model. We observe that the components that make up MELD—bilirubin, INR, and creatinine—are the top three most important variables. Age and years on the waitlist are also highly important. Cancer is less important, perhaps because the model is capable of identifying HCC patients in another way.

Table 4.2 presents the mean number of adverse events that are averted for cancer patients in simulation by each of the three methods over the five sample paths of organ arrivals obtained from previous years. In every region, random forest averts significantly more adverse events than unadjusted MELD. In 6 out of the 11 regions, random forest also outperforms HCC-adjusted MELD.

Figure 4-3 visualizes the information in Tables 4.1 and 4.2. Figure 4-4 demonstrates the percentage improvement for both random forest and unadjusted meld HCC-adjusted MELD for all patients. Figure 4-5 depicts the same comparison for cancer patients. We observe that in almost all of the regions, both random forest and unadjusted MELD outperform the HCC-adjusted MELD when considering all patients. However, for cancer patients, we see that random forest improves upon HCC-adjusted MELD in six regions, while unadjusted MELD drastically underperforms HCC-adjusted MELD in all regions when we restrict our attention to cancer patients.

Based on these results, we observe that HCC-adjusted MELD went too far in
Figure 4-2: Geography of service regions. There are 11 service regions in all, with Alaska and Hawaii (not shown) belonging to region 6. For the most part, livers are procured and allocated in each region in parallel.

<table>
<thead>
<tr>
<th>Region</th>
<th>All Adverse Events</th>
<th>RF Averted</th>
<th>Unadjusted MELD</th>
<th>HCC-MELD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>197</td>
<td>99.8</td>
<td>92.6</td>
<td>59.4</td>
</tr>
<tr>
<td>2</td>
<td>561</td>
<td>349.4</td>
<td>313.0</td>
<td>271.8</td>
</tr>
<tr>
<td>3</td>
<td>291</td>
<td>230.4</td>
<td>225.4</td>
<td>225.0</td>
</tr>
<tr>
<td>4</td>
<td>470</td>
<td>243.6</td>
<td>227.6</td>
<td>180.0</td>
</tr>
<tr>
<td>5</td>
<td>656</td>
<td>369.6</td>
<td>324.4</td>
<td>265.2</td>
</tr>
<tr>
<td>6</td>
<td>93</td>
<td>58.6</td>
<td>54.6</td>
<td>46.8</td>
</tr>
<tr>
<td>7</td>
<td>318</td>
<td>203.0</td>
<td>176.2</td>
<td>149.2</td>
</tr>
<tr>
<td>8</td>
<td>243</td>
<td>169.0</td>
<td>146.0</td>
<td>143.6</td>
</tr>
<tr>
<td>9</td>
<td>285</td>
<td>144.6</td>
<td>134.0</td>
<td>109.6</td>
</tr>
<tr>
<td>10</td>
<td>219</td>
<td>158.2</td>
<td>144.2</td>
<td>144.0</td>
</tr>
<tr>
<td>11</td>
<td>239</td>
<td>193.8</td>
<td>179.2</td>
<td>174.2</td>
</tr>
<tr>
<td>Total</td>
<td>3699</td>
<td>2220.0</td>
<td>2017.2</td>
<td>1768.8</td>
</tr>
</tbody>
</table>

Table 4.1: Mean number of adverse events averted for all patients in simulation by using random forest, unadjusted MELD, and HCC-adjusted MELD for organ arrival sample paths from previous years. Random forest outperforms unadjusted MELD by 10.0% and HCC-adjusted MELD by 25.5% in terms of number of adverse events averted for all patients.
Table 4.2: Mean number of adverse events averted for cancer patients in simulation by using random forest, unadjusted MELD, and HCC-adjusted MELD for organ arrival sample paths from previous years. Random forest outperforms unadjusted MELD by 63.1% and HCC-adjusted MELD by 2.8% in terms of number of adverse events averted for cancer patients.

giving an advantage to HCC patients when compared to unadjusted MELD. In every region, HCC-adjusted MELD underperforms the original unadjusted MELD in number of adverse events averted for all patients in simulation. We observe that the random forest model improves upon the performance of unadjusted MELD with respect to all patients, while still averting at least as many adverse events as HCC-adjusted MELD, indicating that it is not necessary to sacrifice overall performance to ensure an equitable allocation for cancer patients.

4.5 Discussion

Every day, life-or-death decisions must be made by transplant practitioners when deciding which patients on the liver waitlist are in most dire need of transplantation. For almost two decades now, MELD has provided the backbone for the scoring system used to rank patients on the waitlist, with varying levels of success for different patient populations. In particular, HCC-adjusted MELD, the current scoring system, gives an advantage to cancer patients at the expense of the overall patient population.
Figure 4-3: Mean number of adverse events averted by region in simulation by random forest, unadjusted MELD, and HCC-adjusted MELD for organ arrival sample paths from previous years. The solid (transparent) horizontal line in each region denotes the total number of adverse events for cancer (all) patients. The solid (transparent) colored bars in each region denote the number of adverse events averted by each ranking method for cancer (all) patients. In all service regions, random forest dominates unadjusted MELD in number of adverse events averted for both all patients and cancer patients. Unadjusted MELD outperforms HCC-adjusted MELD for all patients, but significantly underperforms for cancer patients. In all service regions, random forest significantly outperforms HCC-adjusted MELD for all patients, while providing comparable performance for cancer patients.

While the original implementation of MELD led to a sharp decline in waitlist mortality compared to the previous ranking metric of time accrued on the waitlist, many problematic issues have arisen. These include: the inability to appropriately capture the materially different disease progression of HCC patients; capturing only waitlist mortality and not probability of becoming unsuitable for transplant; and using the information captured by other important variables such as serum albumin levels, age of
Figure 4-4: Percentage increase in adverse events averted by random forest and unadjusted MELD over HCC-adjusted MELD for each region for organ arrival sample paths from previous years. The error bars denote 95%-confidence intervals. For all service regions, the percentage increase in averted adverse events by random forest over HCC-adjusted MELD is significant.

Figure 4-5: Percentage change in adverse events averted by random forest and unadjusted MELD over HCC-adjusted MELD for cancer patients in each region for organ arrival sample paths from previous years. The error bars denote 95%-confidence intervals. In 6 out of 11 regions, random forest improves upon HCC-adjusted MELD. In all regions, unadjusted MELD performs worse than HCC-adjusted MELD for cancer patients.
patient, years accrued on the waitlist, and the trajectory of patients’ clinical scores. As the transplant community has become aware of each of these problems, incremental modifications have been proposed and sometimes implemented as part of the allocation policy. However, these tweaks have often been justified via intuition and incorporated as part of current practice before thorough scientific validation. Only once retrospective studies have been conducted and after additional patients’ lives placed at risk are additional changes to current policy considered.

We developed an analytical tool that takes all available patient information to make a prediction as to whether or not the patient will undergo the adverse events of either death or becoming unsuitable for transplant within three months. In contrast to the piecemeal way in which current policy has been constructed, our tool is trained on historical outcomes in a unified fashion: instead of adding in exceptions and cutoffs ex-post in an attempt to decrease mortality on the waitlist, our methods tackle the problem directly by building these different criteria into the model itself. Indeed, our simulation studies demonstrate that although the addition of these arbitrary cutoffs do in fact help HCC patients, this additional benefit comes at a cost for the patient population as a whole.

While it is the case that the components of the MELD score (bilirubin, INR, and creatinine) are highly effective predictors (in fact, the most important variables as we see in Figure 4-1), they are by no means the only relevant predictors. As we observed in our simulation, our model improves upon both unadjusted and HCC-adjusted MELD in terms of adverse events averted for the entire patient population by 10-25%. Importantly, the added benefit to the overall patient population does not require a tradeoff from cancer patients; our model provides comparable performance for cancer patients to the MELD policy that was intentionally adjusted to favor cancer patients.

The random forest model significantly outperforms the current scoring policy in use, HCC-adjusted MELD, in averting adverse events for all patients. It is important to note here that our simulation procedure is a close approximation to real-world dynamics. Although we censor patient observations where we cannot observe the
patient’s true three-month mortality outcome due to transplantation, our simulation does capture all patients who were in fact on the liver waitlist at those particular points in time, as well as their subsequent outcomes. Thus, by simulating the allocation policies of both MELD scores and the random forest ranking, we are observing the number of patients who would have been saved by transplantation if additional organs (above the existing procurement rate) were procured, without having to resort to counterfactual conjectures of patient survivability. Put another way, our simulation experiments readily inform us of the impact that an increase in the donation rate of deceased-donor livers would have in decreasing the rate of adverse events for patients with end-stage liver disease.

Application of methods from the predictive analytics and machine learning literature have enabled the development of a more accurate model for the problem of predicting adverse events within three months for patients on the liver transplantation waitlist. We have shown that our random forest model not only improves on out-of-sample AUC when compared to the existing ways of computing MELD scores, but also yields a significant improvement in the number of lives saved (and adverse events averted) via transplantation. Importantly, our model provides actionable insight into implementing a more fair allocation policy without unnecessary tradeoffs for all patients, regardless of disease etiology.
Chapter 5

Conclusion

At the beginning of this thesis, we aspired to address three important problems plaguing the US health care system today: the uncontrolled increases in expenditures for health care procedures; the acceptance and rejection decisions of kidneys offered to waitlisted patients; and decreasing overall waitlist mortality while improving fairness for patients on the liver transplantation waitlist. In each of these problem settings, we presented solution approaches and frameworks based on modern data analytics methods that can take as input existing datasets and produce prescriptive insights and actionable decisions that yield significant benefits over current practice.

To tackle the ever-rising health care costs that have been troubling employers, we presented a framework driven by an observation readily borne out in the data: the cost of the same health care procedure can often vary dramatically—in some cases, by orders of magnitude—between providers that are located within close geographical proximity. In the absence of incentives to influence behavior regarding provider choice, patients may by chance seek to obtain medical procedures by providers who charge excessive amounts. We proposed an approach starting with historical insurance claims data that identifies all relevant data. Our approach uses mixed integer linear optimization to help employers identify prospective partner providers in the provision of medical procedures that adequately covers the patient population at low cost. Our approach leads conservatively to a 10% reduction in health care costs in a real-world study of a private employer with more than 14,000 dependents.
In an effort to move beyond irreproducible decision-making on the part of transplant practitioners who must make the vital decision of whether or not to accept or reject an offered kidney on behalf of a patient on the kidney transplant waitlist, we developed a data-driven random forest model to accurately predict the probability that a kidney of desired quality will be offered to a particular patient within three, six, or twelve months. Our model accounts for a patient’s OPO, blood group, wait time, DR antigens, and prior offer history to provide accurate and personalized predictions. These predictions hold potential to facilitate timely transplantation and optimize the efficiency of the allocation overall.

Current policy in allocating deceased-donor livers to patients on the liver transplant waitlist has largely been shaped by trial-and-error since the initial introduction of the MELD score. Haphazard and often arbitrary heuristic modifications to the original implementation of MELD have led to the undesirable tradeoff between improving overall patient outcomes and ensuring fairness between all patient subpopulations. To this day, the effects of these tweaks to the MELD score are not well understood, and, as a result, additional patient lives are placed in danger. We proposed a data-driven model that not only improves upon the out-of-sample AUC when predicting three-month likelihood of experiencing an adverse event for liver waitlist patients, but also leads to a 25% improvement in the number of deaths averted over the current HCC-adjusted MELD scoring policy in simulation. Importantly, this improvement does not come at the expense of cancer patients. The predictions of our model could potentially form the basis for a new, unified scoring metric on which to rank the severity of patients with end-stage liver disease.

For each of these three problems, we detailed the entire process whereby we transform raw data and distill it into prescriptive, actionable insights that can lead to a material difference in terms of both reducing costs and improving patient outcomes. These three problem settings serve to illuminate just a few of the areas where data analytics can yield an edge over current practice; the time is ripe for modern techniques in data analytics to transform the US health care system more broadly.
Appendix A

Additional Figures

- A-1 Histogram of price changes for Upper GI Endoscopy Biopsy
- A-2 Histogram of travel distances for Upper GI Endoscopy Biopsy
- A-3 Histogram of price changes for CT Scan of Abdomen and Pelvis
- A-4 Histogram of travel distances for CT Scan of Abdomen and Pelvis
- A-5 Histogram of price changes for MRI of Brain
- A-6 Histogram of travel distances for MRI of Brain
- A-7 Histogram of price changes for Lesion Removal Colonoscopy
- A-8 Histogram of travel distances for Lesion Removal Colonoscopy
Figure A-1: Histogram of price changes for Upper GI Endoscopy Biopsy.

Figure A-2: Histogram of travel distances for Upper GI Endoscopy Biopsy.
Histograms of Pricing for CT ABD & PELV W/CONTRAST

Figure A-3: Histogram of price changes for CT Scan of Abdomen and Pelvis.

Histogram of Distances Traveled for CT ABD & PELV W/CONTRAST

Figure A-4: Histogram of travel distances for CT Scan of Abdomen and Pelvis.
Histograms of Pricing for MRI BRAIN W/O & W/DYE

![Histogram of price changes for MRI of Brain.](image)

Figure A-5: Histogram of price changes for MRI of Brain.

Histogram of Distances Traveled for MRI BRAIN W/O & W/DYE

![Histogram of travel distances for MRI of Brain.](image)

Figure A-6: Histogram of travel distances for MRI of Brain.
Histograms of Pricing for LESION REMOVAL COLONOSCOPY

Figure A-7: Histogram of price changes for Lesion Removal Colonoscopy.

Histogram of Distances Traveled for LESION REMOVAL COLONOSCOPY

Figure A-8: Histogram of travel distances for Lesion Removal Colonoscopy.
Bibliography


